Haaga-Helia
ammattikorkeakoulu Oy

Datavisualisoinnin hyödyntäminen

Ari Nieminen

Opinnäytetyö
Tietojenkäsittelyn koulutusohjelma
2018

Haaga-Helia ammattikorkeakoulu Oy

Tiivistelmä

14.5.2018

Tekijä(t)
Ari Nieminen

Koulutusohjelma
Tietojenkäsittelyn koulutusohjelma

Opinnäytetyön otsikko
Datavisualisoinnin hyödyntäminen

Sivu- ja liitesivumäärä
33 + 1

Opinnäytetyön otsikko englanniksi
Utilizing data visualization

Tämä opinnäytetyö käsittelee datavisualisoinnin hyödyntämistä analysoitaessa datasta jalostettua informaatiota. Tavoitteena on antaa lukijalle kuvaus datavisualisoinnin mahdollistamista keinoista sekä ymmärtää dataa paremmin ja löytämään helpommin eroavaisuuksia tutkittavan aineiston sisällä. Opinnäytetyöllä ei ole toimeksiantajaa. Työ jakautuu teoriaosuuteen, empiiriseen osaan sekä pohdintaan.

Teoriaosuudessa tarkastellaan tiedon visualisointia hahmolakien sekä havaintopsykologian näkökulmasta sekä ihmisen luontaista kykyä hahmottaa kuvia, värejä ja kokoeroja paremmin kuin pelkkää tekstiä ja numeroita.

Empiirisessä osuudessa tarkastellaan datavisualisoinnin hyödyntämisen mahdollistavia tekijöitä, eli mitä kaikkea datalle pitää ensin tehdä jotta siitä syntyisi informaatiota, jota voisi sen jälkeen datavisualisoinnilla hyödyntää. Lisäksi käyn läpi tiedon itsepalveluvisualisointivälineiden markkinoita sekä suosituimpia ohjelmistoja. Tarkastelen case-tyyppisen toimeksiannon kautta avoimen datan hyödyntämistä ja tiedon visualisointia Tableau Desktop –visualisointiohjelmistolla.

Opinnäytetyön tutkimuksen tuloksina voidaan todeta, että tiedon määrän eksponentiaalisen kasvun myötä on myös tiedon hyödyntäminen muuttunut haasteellisemmaksi. Tiedon siiloutuminen sekä hajauttaminen perinteisiin konesaleihin sekä pilvipalveluihin on tuonut haasteita kasvaneen tietomäärän hallintaan ja hyödyntämiseen. Big datan myötä käytettävissä olevan datan määrä on kasvanut myös yritysten ulkopuolella ja näin ollen käsittelemätöntä dataa sekä osittain jalostettua dataa on yhä moninaisemmassa muodossa tarjolla hyödynnettäväksi myös globaalisti. Yritysten strategisen päätöksenteon tukemiseksi tarvitaan analysointia olemassa olevasta datasta nopeasti ja tehokkaasti. Tiedon visualisointi on noussut merkittävään asemaan vastattaessa tähän haasteeseen.

Asiasanat
visualisointi, raportointi, analytiikka, tiedonhallinta, havaintopsykologia

Sisällys

Käsitteet

Big data	‘ei rakenteellista’ massadataa
Data mart	tieto-varastosta/lähteestä jalostettu sisällöltään spesifinen tietomalli tai kuutio
Iot	esineiden internet, teollinen internet (Internet of Things)
ETL prosessi	tiedonsiirtoprosessi lähdejärjestelmistä tietovarastoon (extract transform load)
Data	merkkejä ja symboleja sisältävää digitaalisesti tallennettua potentiaalista informaatiota
Informaatio	datasta tulkittua, muotoiltua tai muodostettua tietoa
Tieto	informaatiosta tehtyjä havaintoja päätöksenteon tueksi
Strukturoimaton data	ei rakenteellista, volyymiltaan suurta määrää dataa
Strukturoitu data	edellisestä louhittua rakenteellista (hierarkista) dataa
Niche	pienen markkina-alueen segmentti
Päätöspuu -malli	analyysimalli päätösvaihtoehtojen ja seurausten hahmottamiseen
Spektri	valon aallonpituusalueen värijakauma
ODBC	avoin tietokantarajapinta

1 Johdanto

Visualisointi ei ole mikään uusi ilmiö. Kuten Väisänen (2017 15.8.2017) kirjoituksessaan toteaa, on visualisointeja tehty jo 1700-luvulta asti, kuvaten esimerkiksi Napoleonin sotaretkiä, epidemioiden leviämistä sekä väestönkehityksen trendejä. Mikä sitten datan visualisoinnin puolella on niin ihmeellistä? Väisäsen (2017 15.8.2017) mukaan se on interaktiivisten visualisointiohjelmistojen kasvanut kysyntä.

Tiedon määrän kasvaessa eksponentiaalisesti tuo se haasteita tiedon hyödyntämisrajapintaan. On ennustettu että datan kokonaismäärä kaksinkertaistuu joka toinen vuosi ja vuonna 2020 datan määrä olisi n. 50-kertainen ja palvelinten määrä kasvaisi kymmenkertaiseksi nykyhetkeen verrattuna (Taloussanomat 2011.)

Interaktiiviset datan itsepalveluohjelmistot, kuten QlikSense, Tableau ja Power BI ovat tuoneet datavisualisoinnin suuren yleisön tietoisuuteen. Perinteiset Excel- sekä muut vastaavanlaiset taulukkolaskentaohjelmat ovat kykenemättömiä suoriutumaan datan käsittelystä nykyvaatimusten mukaisesti. Interaktiivinen itsepalveluvisualisointi antaa käyttäjälle mahdollisuuden tutkia dataa itsenäisesti ja tehdä omia johtopäätöksiä havainnoista (Väisänen 2017 15.8.2017.)

Miten analysoida ja löytää suuresta tietomassasta nopeasti ja tehokkaasti hyödyllinen tieto. Datasta sovelletut graafiset esitykset kertovat korrelaatioista eli muuttujien välisistä riippuvuussuhteista ja poikkeamista paljon nopeammin kuin numeroita ja tekstiä sisältävät taulukot. Tiedon visualisoinnin avulla käyttäjät voivat havaita nämä korrelaatiot ja poikkeamat aineistosta yhdellä silmäyksellä ja ryhtyä näin ollen nopeammin asianmukaisiin toimiin (Eckerson, W. & Hammond, M. 2011.)

Yhdistettäessä dataa moninaisista lähteistä ja huomioiden sen kiihtyvän kasvun, ollaan ongelman ytimessä. Dataa pitää siirtää, tallentaa, muokata ja yhdistellä sekä analysoida mahdollisimman tehokkaasti, jotta sitä voisi myös hyödyntää (Salo 2013, 21.) Ihmisten tuottaman datan lisäksi automaattista dataa syntyy erilaisista mittauslaitteista, valvontalaitteista, huoltojärjestelmistä, terveydenhuollon järjestelmistä, älypuhelimista, sensoreista ja monista muista tietoverkkoihin kytketyistä laitteista. Kaikkea edellä mainittua dataa toki hyödynnetään jo nyt omilla sovellusalustoillaan, mutta suuri osa datasta jää hyödyntämättä sekä siitä johdetusta informaatiosta tallentamatta (Salo 2013, 21.)

Tiedon analysointia varten ovat raportointiohjelmistojen markkinoita hallinneet datan visualisointiin erikoistuneet itsepalvelutyökalut, jotka helppokäyttöisyydellään auttavat hyödyn-

täjää ymmärtämään datan sisältöä nopeasti ja vaivattomasti. Tiedon visualisoinnilla voidaan paremmin ymmärtää mitä data voi meille kertoa. Näiden ohjelmistojen käyttäjäkynnys on madaltunut niiden helppokäyttöisyyden myötä, kuka tahansa yrityksen liiketoimintayksikön edustaja voi tehdä havaintoja käytettävissä olevista tietomalleista itsenäisesti ilman yrityksen IT-osaston apua. IT-tukitoiminnoilta ei vaadita kuin toiminnan mahdollistaminen. Toki tietomallien rakentamisella ymmärrettävään muotoon sisällön osalta, on iso merkitys hyödyntäjän näkökulmasta.

Olen työskennellyt yli kymmenen vuotta tiedolla johtamisen parissa ja toteuttanut perinteistä taulukkoraportointia sekä datavisualisointeja eri raportointi- ja analytiikkavälineillä. Olen ollut mukana myös toteuttamassa tietovarastohankkeita, joissa eri lähdejärjestelmistä kerättyä dataa on johdettu ETL-prosessin läpi tietovarastoihin. Olen viimeisen kymmenen vuoden aikana havainnut miten BI-analytiikan itsepalvelutyökalujen sekä perinteisempien suurien BI-alan toimittajien raportointi- ja analytiikkavälineiden yhteinen kehityssuuntaus on painottunut yhä voimakkaammin interaktiivisuuden, visuaalisuuden sekä helppokäyttöisyyden korostamiseen. Käyttäjäkohderyhmänä ovat nyt johtajat ja päätöksentekoon vaikuttavat henkilöt, eivät enää perinteiset IT-osastot.

1.1 Tutkimusongelma

Digitaalisen tiedon määrä on kasvanut huimaa vauhtia sosiaalisen median, big datan ja Iot:n myötä. Lisäksi yritysten tallentama digitaalinen data erilaisista liiketoimintatapahtumista on lisääntynyt. Tästä suuresta tietomassasta pyritään tallentamaan kaikki se oleellinen informaatio jota voidaan hyödyntää päätöksenteon tukemiseksi. Informaatiosta pitää jalostaa tietoa ja esittää se ymmärrettävässä muodossa. Tätä tietoa analysoidaan ja raportoidaan erilaisilla raportointiin ja analytiikkaan erikoistuneilla interaktiivisilla tiedon visualisointivälineillä.

Opinnäytetyössä selvitetään seuraavien tutkimuskysymysten kautta:

- Miksi hyödyntää datavisualisointia perinteisen taulukkoraportoinnin sijaan?
- Miten tietoa kannattaa visualisoida?

Tiedon visualisoinnilla on merkittävä rooli tiedolla johtamisen mahdollistajana. Sen avulla voidaan, aineiston kokomäärästä riippumatta, tuoda esiin monimutkaistenkin ongelmien taustalla piilevät syyt joita olisi perinteisillä taulukkolaskentamallisilla näkymillä vaikea havaita. Kaikkea tietoa ei tietenkään pidä visualisoida, vaan löytää paras mahdollinen vaihtoehto ja kombinaatio eri tilanteisiin. Joskus on hyvä pidättäytyä perinteisessä taulukkomallisessa esitystavassa, mutta siihenkin on mahdollista ja suositeltavaa lisätä visuaalisia

elementtejä nostamaan oleellinen osa tiedosta korostettuna esiin. Tästä esitän konkreettisen esimerkin opinnäytetyön empiirisen osion case -osuudessa.

Miten sitten toteutetaan datavisualisointia oikein? Onko olemassa oikeaa ja väärää tapaa? Visualisointi vaatii suunnittelua, resursseja sekä osaamista. Tekijän on ymmärrettävä dataa myös isommassa kuvassa ja saada datasta tuotua esiin kaikkein olennaisin tieto. Tiedolla johtaminen on tullut tärkeäksi käsitteeksi yritysten pyrkiessä kasvattamaan liiketoimintaansa kilpailijoita tehokkaammin. Se kuka kykenee hyödyntämään parhaiten omat ja ulkoiset tietolähteet, hyödyntäen edistyneen analytiikan avulla tuotettuja ennusteita tulevaisuuden skenaarioista, on etulyöntiasemassa markkinoilla. Tähän kaikkeen tarvitaan datavisualisointia avuksi. Opinnäytetyön tarkoituksena on antaa lukijalle hyvän käytännön mukaiset perusteet datavisualisoinnin hyödyntämiseen, huomioiden psykologiset tekijät hahmolakien osalta, sekä otannan nykypäivän BI-analytiikan ja raportoinnin interaktiivisista työkaluista.

1.2 Viitekehys

Työni keskittyy käsittelemään teoriaosuudessa visualisoinnin etuja hahmopsykologian kautta, perustuen ihmisen luontaiseen kykyyn hahmottaa kuvioiden ja värien symboliikkaa paremmin kuin pelkkää tekstiä ja numeroita. Empiirinen osuus (luku 3) perustuu havainnoimaan keskitetyn tietovaraston merkitystä tiedon johtamisen näkökulmasta sekä miten kasvavasta tietomassasta jalostetaan rakenteellisia tietolähteitä visuaalisten raportointityökalujen käyttöön. Käsittelen opinnäytetyössäni myös big data käsitettä eksponentiaalisesti kasvavan tiedon määrän kuvaajana, mutta en käsittele big dataa arkkitehtuurinäkökulmasta. En myöskään käsittele tiedon louhintaa sekä koneoppimista teknisestä näkökulmasta, vaan yhtenä tiedon jalostuksen prosessina.

Opinnäytetyön lopputuloksena on kattava katsaus tiedon visualisoinnin merkityksestä datassa piilevän informaation oivaltamiseksi, huomioiden havaintopsykologiset tekijät sekä visualisoinnin mahdollistavat itsepalveluraportoinnin välineet.

1.3 Tutkimusmenetelmät

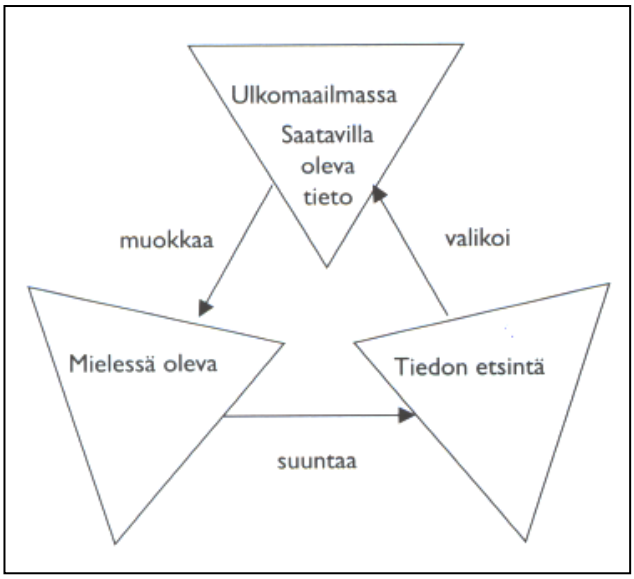
Opinnäytetyön tutkimusmenetelmänä on kirjallisuuskatsaus ja oman työkokemuksen tuoma näkemys kasvaneeseen tarpeeseen ymmärtää datasta luodun informaation sisältöä visualisoinnin keinoin. Tutkin havaitsemisen teorioita havaintopsykologian näkökulmasta, miten ihminen havaitsee poikkeavuudet ja korrelaatiot tehokkaammin, kun dataa on esitettynä visuaalisesti. Käsittelen datavisualisointiin tarkoitettujen ohjelmistojen ominaisuuksia

ja alan markkinoita Gartnerin nelikenttäanalyysin kautta. Käyn läpi tietovarastoinnin perusteita ja ETL-prosessia sekä tiedon jalostamisen merkitystä raportoinnin ja analytiikan taustalla. Toteutan case-luonteisen datavisualisointi toimeksiannon Tableau -ohjelmistolla kotimaisten yritysten pilvipalveluiden käyttöönoton kehityksestä vuosina 2013 - 2017, hyödyntäen Tilastokeskuksen julkaisemaa avointa dataa.

Opinnäytetyön aiheen ajankohtaisuuden sekä aihealueen johdosta käytän tässä työssä paljon kuvamateriaalia, jotta datavisualisoinnin merkityksen sanoma tulisi esiin mahdollisimman tehokkaasti. Käytän myös paljon Business Intelligence eli BI-alan blogi- ja verkkosivuja lähteinä, koska datavisualisoinnin ja analytiikan kenttä on muuttunut viime vuosina nopeasti tiedon määrän lisääntymisen myötä sekä tiedolla johtamisen noustessa yhä suurempaan rooliin tukemaan päätöksentekoa. Tästä syystä tuorein tieto alan trendeistä, visualisointivälineiden päivitetyistä ominaisuuksista sekä BI -alan markkinoista, ei löydy kirjajulkaisuista, vaan globaalista Internetistä.

2 Havaitsemisen teorioita

Ulric Neisserin (1982, 10) mukaan kognitiolla tarkoitetaan tiedon hankintaa, sen järjestämistä ja käyttöä, jotka yhdistyvät tietämisen toiminnaksi. Kognitiivisen psykologian kirjoissa käsitelläänkin ensin havaitsemista, ja vasta sen jälkeen muistia sekä muita tiedon sisäistämisen toimintoja (Neisser 1982, 19). Ennakoivat skeemat ovat näköaistin kannalta ratkaisevia kognitiivisia rakenteita, joiden kautta havaitsija vastaanottaessa näköhavaintoa, hyväksyy toisenlaista tietoa muun totutun tiedon sijaan (Neisser 1982, 24). Ennakoivat skeemat ovat havaintotoimintaan liittyviä suunnitelmia ja valmiuksia vastaanottamaan määritellyin tavoin rakennettua optista tietoa, kuten Neisser (1982, 25) asian kuvaa havaintosykliä esittävässä kuvassa.



Kuva 1. Havaintosykli (Neisser, U. 1982, 25)

Sinkkonen, Kuoppala, Parkkinen & Vastamäki (2006, 71) toteavat, että fysikaaliset kohteet sekä ulkomaailmasta heijastuvat valonlähteet säteilevät tai heijastavat valon eri aallonpituuksia. Niiden kohdatessa silmän verkkokalvon, lähettävät silmän aistinsolut ärsykkeen saatuaan signaaleja edelleen aivoihimme. Aivokuoremme hermosoluissa tapahtuu väri-informaation käsittelyä, ääriviivojen suuntien käsittelyä sekä liikkeen havainnointia, joiden aistimuksista aivomme tulkitsevat havaintokuvan (Sinkkonen ym. 2006, 71.)

Näköaistin kohdatessa yksittäisiä ärsykkeitä havaintojärjestelmämme ryhmittelee ne isompiin kokonaisuuksiin, joista havaittu kohde pyritään tunnistamaan hyödyntäen mielemme odotuksia sekä aiempien havaintojen kautta tutuksi tulleita luokitteluja. Kaikista ärsykkeiden yksityiskohdista ja yhdistelmistä ihminen hahmottaa yleisimmin ne jotka ovat yksinkertaisia ja tuttuja (Sinkkonen ym. 2006, 89.) Erillään olevia kuva-alkioita yhdistele-

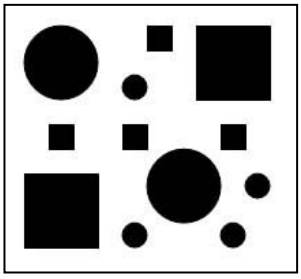
mällä ne koetaan saman kohteen osiksi tai kokonaisuuksiksi. Kun aivomme tekevät tämänkaltaista ryhmittelyä toistuvasti, niin näistä osista muodostuu yhä laajempia kokonaisuuksia (Sinkkonen ym. 2006, 89.) Näitä ihmisen synnynnäisiä piirteiden yhdistelytapoja, kuvaavat parhaiten hahmolait (Sinkkonen ym. 2006, 89).

2.1 Gestaltin hahmolait (Gestalt laws)

Laine (18.2.2004) kuvailee tutkielmassaan Gestaltin hahmolakeja. Sana ” Gestalt” tulee saksan kielestä ja tarkoittaa ”hahmoa”. Saksassa ja Itävallassa 1890 luvulla syntynyt Gestalt -teoria, jonka taustalla oli hahmopsykologinen koulukuntasuuntaus. Tämän suuntauksen kohteena oli tutkia, kuinka mielemme muodostavat kokonaisuuksia nähdessämme epätäydellisiä ryhmittelyjä. Teoria hahmolaeista pyrkii osoittamaan tavan, jolla aivot muodostavat kokonaisuuksien yhdistelmiä eri visuaalisten havaintojen yksityiskohdista. Yleisimmät ja käytetyimmät Gestaltin hahmolait ovat kokonaisuuden laki (engl. completeness) sekä tunnettuuden laki (engl. familiarity). Ihminen hahmottaa yleensä ensin kokonaisuuden josta itselleen tutut kuviot hahmottuvat ensimmäisinä. (Laine 18.2.2004.) Gestaltin hahmolaeista voidaan datan visualisoinnin näkökulmasta tarkastella Laineen (18.2.2004) mukaan seuraavia ryhmittelyyn perustuvia lakeja, jotka ovat tarkemmin esitelty alaluvuissa 2.1.1, 2.1.2 ja 2.1.3.

2.1.1 Samankaltaisuus (engl. similarity)

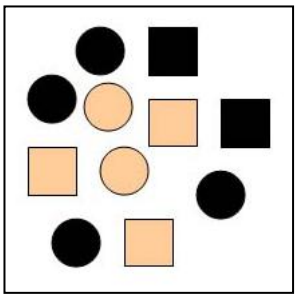
Ihminen mieltää samankaltaiset muodot, värit sekä kuviot yhteenkuuluviksi. Mitä enemmän kohteet ovat toistensa näköisiä, aivot mieltävät niiden muodostavan ryhmiä. Samankaltaisuuden lakia hyödynnetään esimerkiksi ”Scatter Plot” mallinnuksessa.



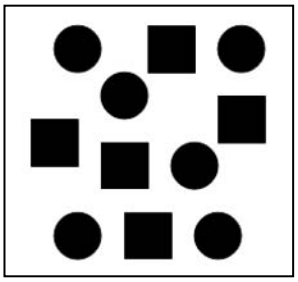
Kuva 2. Erikokoiset, erimuotoiset ja samanväriset objektit (Laine 18.2.2004)

Yllä olevassa kuvassa on objektit esitetty samanvärisinä mutta erikokoisina ja erimuotoisina. Yleisimmin katsojan kohde kiinnittyy objektien kokoeroon enemmän kuin muotojen eroon ja näin ollen kokoeroavaisuutta käytetään usein ryhmittelyyn (Laine 18.2.2004.)

Kuvassa 3 ovat objektit esitetty samankokoisina, mutta väriltään erilaisilta. Kun värisävyt ovat huomattavasti toisistaan poikkeavat, voidaan väriä käyttää onnistuneesti ryhmittelyssä. On kuitenkin huomioitava, ettei käytä esimerkiksi harmaan eri sävyjä, joita on hankala silmällä erottaa. (Laine 18.2.2004.)



Kuva 3. Erimuotoiset ja eriväriset objektit (Laine 18.2.2004)

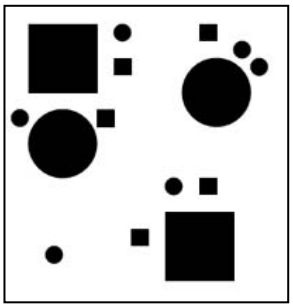


Kuva 4. Erimuotoiset ja samanväriset objektit (Laine 18.2.2004)

Kuvassa 4 kiinnittyy huomio vain objektien muotoon. Se on ainoa toisistaan erottava tekijä. Tämä esitystapa sopii hyvin, jos aiempia ryhmittelyesimerkkejä ei voida käyttää. Kun muotoerot ovat selkeitä ja yksinkertaisia, kuten kuvassa, niiden ryhmittely on ihmiselle luontaista (Laine 18.2.2004.)

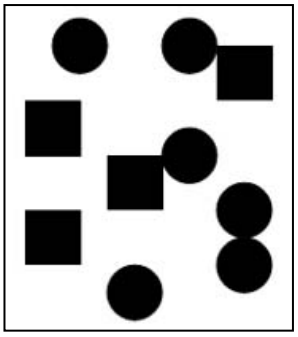
2.1.2 Läheisyys (engl. proximity)

Kun kuviot on sijoiteltu lähekkäin, ne mielletään yhteenkuuluviksi. Läheisyyslaki perustuu etäisyyssuhteisiin, eli mitä lähempänä objektit sijaitsevat toisiaan, sitä todennäköisemmin ne tiedostetaan ryhmäksi.



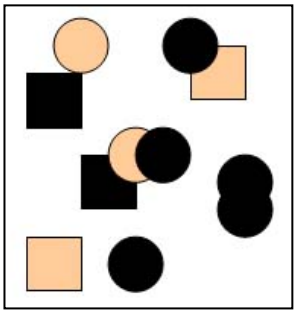
Kuva 5. Lähellä sijaitsevat objektit (Laine 18.2.2004)

Yllä näkyvässä kuvassa (kuva 5, 9) on kuvattu kolme ryhmää ja yksi ryhmään kuulumaton objekti, pikkuympyrä alakulmassa. Ryhmittelyssä on käytetty kokoeroa sekä muotoa sekaisin, eikä se ole enää aivan yhtä selkeästi eroteltavissa kuin aiemmat esimerkit. Tämänkaltaista ryhmittelyä käytetään kuitenkin usein tekstinkäsittelyssä (Laine 18.2.2004.)



Kuva 6. Kosketuksessa olevat objektit (Laine 18.2.2004)

Kuvassa 6 on osa objekteista kosketuksissa toisiinsa ja ne muodostavat havaittavia siteitä toisiinsa ryhmän sisällä. Huomattavaa on kahden samanmuotoisen objekti kosketus, joka erottuu joukosta parhaiten, koska se perustuu myös samankaltaisuuden lakiin. (Laine 18.2.2004.)

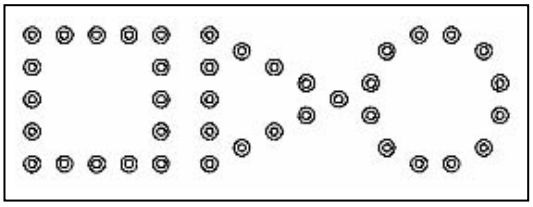


Kuva 7. Limittäin olevat objektit (Laine 18.2.2004)

Kuvassa 7 erimuotoiset ja eriväriset, limittäinen sijoittuneet objektit, luovat illuusion syvyydestä. Erot ryhmittelyn sisällä ovat huomattavissa ja vahvimmin erottuvat kaksiväriset limittyneet ryhmät (Laine 18.2.2004.)

2.1.3 Valiomuotoisuus (engl. good shape)

Periaatteena on, että ihminen pyrkii ymmärtämään erilaisia kuvioita parhaiten, jos ne on esitetty mahdollisimman yksinkertaisessa ja säännönmukaisessa muodossa.



Kuva 8. Valiomuotoiset objektit (Laine 18.2.2004)

Kuvassa kahdeksan esitetään objektit ihmiselle tuttuina, yksinkertaisina ja säännönmukaisina kuvioina. Nämä objektit mielletään neliöksi, kolmioksi ja ympyräksi vaikka ne koostuvat kahdesta sisäkkäisestä ympyrästä (Laine 18.2.2004.)

Datavisualisoinnin onnistuneeseen lopputulokseen pääsemiseksi, on Kosken (2015) mukaan syytä muistaa seuraavat tekijät:

- Ole tietoinen mitä, miksi ja kenelle olet tekemässä visualisointia.
- Tutustu esimerkkeihin epäonnistuneista visualisoinneista.
- Käy läpi visuaalisen suunnittelun perusteita.
- Muista hahmolakien merkitys visualisoinnin asettelussa.
- Älä lisää liikaa tarkoituksetonta grafiikkaa.
- Ei enempää kuin kahta kirjasintyyppiä.
- Värejä ei pidä käyttää perusteettomasti.
- Liian värikäs lopputulos vain rasittaa silmiä.
- Visualisointia ei pidä jakaa liian moneen näkymään.
- Vältä ympyrädiagrammin käyttöä, käytä mieluummin pylväsdiagrammia.

Sinkkosen ym. (2006, 97) mukaan havaitsemisen toiminto itsessään ei ole ihmiselle samalla lailla luontaista kuin näkö-, kuulo-, haju- tai makuaistiminen. Tämä johtuu siitä, että ihminen ei pysty kiinnittämään huomioita suurempaan määrään tietoa, kuin mitä hän pystyy aivoissa prosessoimaan. Ihminen ei voi prosessoida yksityiskohtia kuin yhdestä kohteesta kerrallaan, joten rajallinen tietojenkäsittelykapasiteetin käyttö pakottaa toistuvaan aisti-informaation valikointiin (Sinkkonen ym. 2006, 97.)

3 Datavisualisoinnin toteuttaminen massadatasta

Massadatalla tarkoitetaan rakenteetonta ja monimuotoista, eri lähteistä koottua dataa, jota voivat olla esimerkiksi tekstimuotoinen data, kuva- ja musiikkitiedostot. Tämänkaltaisen rakenteettoman massadatan käsittelyssä on isoin haaste saada se esikäsiteltyä rakenteellisempaan muotoon, jolloin siitä on mahdollista tehdä analysointeja hyödyntäen hajautettua laskentaa, kuten esimerkiksi Spark -ohjelmistoa, joka hajauttaa datan laskennan eri laskentaklustereihin (Aunimo 25.10.2017.)

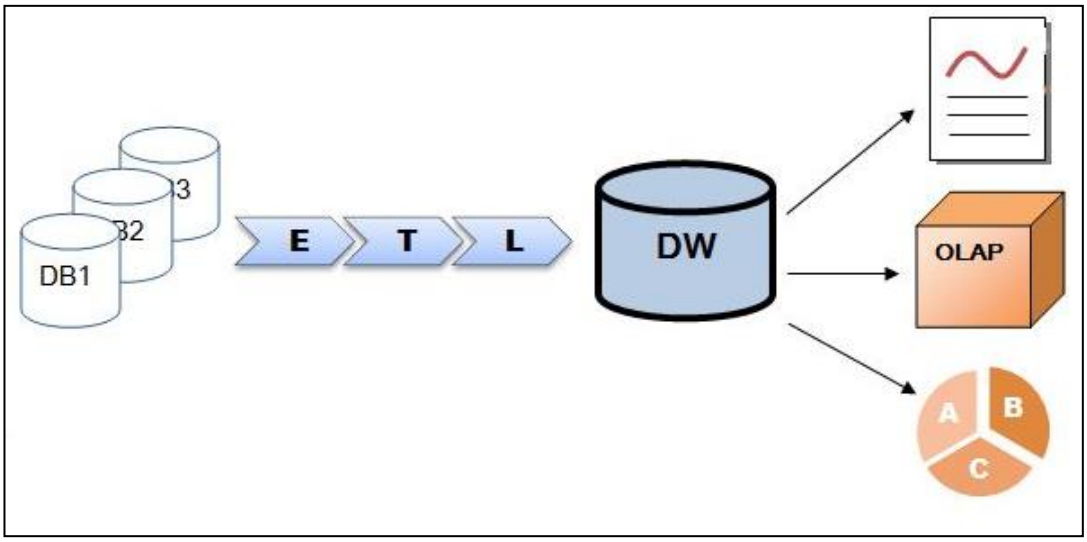
Kanervan (11.9.2016) mukaan datavisualisoinnilla tarkoitetaan digitaalisesta datasta jalostettua tietoa esitettynä visuaalisessa muodossa. Se voidaan nähdä apuvälineenä datan ymmärtämisessä ja sisällön esiintuomisessa. Perinteisestä transaktio- ja masterdatasta voidaan toki tehdä yksinkertaisia laskenta- ja ristiintaulukointia mutta kun datasta jalostettua informaatiota halutaan todella analysoida, ei enää puhuta perinteisestä raportoinnista, vaan analytiikasta. Suurempien tietomassojen, kuten big datan kohdalla, on välttämätöntä hyödyntää visualisointia, jotta tieto olisi ymmärrettävässä muodossa (Kanerva 11.9.2016.)

3.1 Keskitetty tietovarastoratkaisu

Yrityksen liiketoiminta tuottaa yhä enenevässä määrin tallennettavaa tietoa. Monissa suurissa yrityksissä eri liiketoimintayksiköiden ydintiedot ovat yhä tallennettuina eri tietokantoihin, eli yrityksen liiketoimintatietoa on siiloutuneena eri järjestelmiin. Toteutusteknologioita voi olla käytössä useita, osa tiedoista voi olla tallennettuna Oraclen, DB2:n, MySQL:n tai SQL Serverin kantaan, lisäksi yrityksen dataa voi olla tallennettuna myös pilvialustoille, esim. Microsoftin Azure -tallennuspalveluihin. Tämäntapainen hajautettu ydintiedon hallinta ei välttämättä palvele yrityksen päätöksentekoa. Hajautetusta ydintiedosta on vaikeaa ja hidasta tuottaa nopeasti koko yrityksen toiminnan kattavia analyysejä, joista olisi todellista hyötyä tukemaan yrityksen päätöksentekoa.

Yksi ratkaisu nopeampaan ja tehokkaampaan tiedonhallintaan sekä tiedolla johtamiseen on keskitetty tietovarastoratkaisu, jossa yrityksen toiminnalle sekä strategiselle ohjaukselle tärkeä transaktio- ja ydintieto kootaan eri lähdejärjestelmistä yhteiseen tietovarastoon ETL -prosessin avulla. Myös ulkoisista lähteistä voidaan tuoda tietoa yrityksen tietovarastoon. Tietovarastosta mallinnetuista datamarteista tai OLAP-kuutioista koostettujen raporttien, ennusteiden sekä analyysien avulla voivat eri liiketoimintayksiköt hyödyntää yrityksen yhteistä tietovarastosisältöä päätöksentekonsa tueksi.

Yksinkertaistettu tietovarastointiprosessi, jossa eri tietolähteistä kerätty data muokataan ETL- vaiheessa ennen tietovarastoon (DW) lataamista ja edelleen hyödyntämisrajapintoihin, on kuvattu alla. Prosessi etenee vasemmalta oikealle.



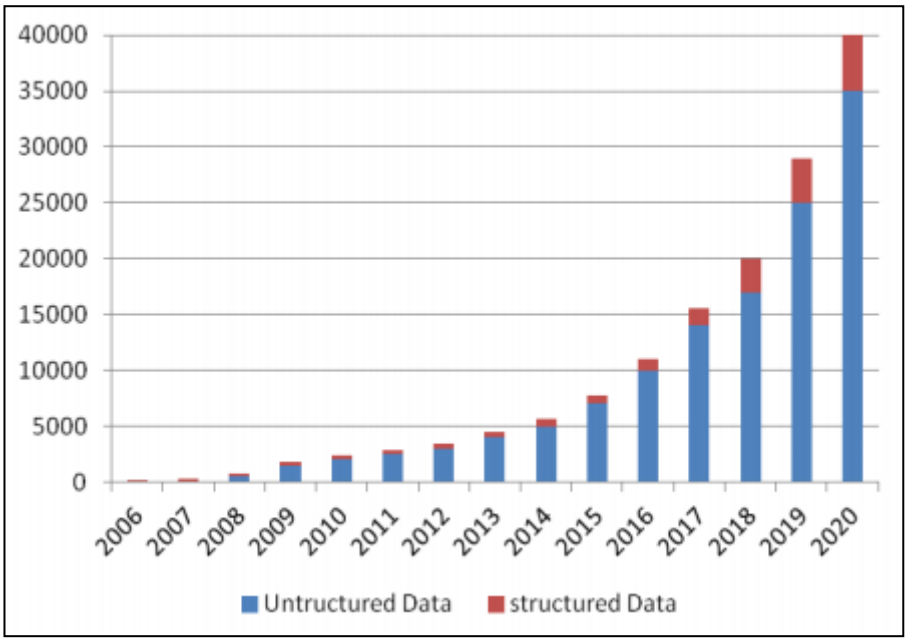
Kuva 9. ETL-prosessin kuvaus (Tietokaira 2015)

3.2 Big data

Big data käsitteellä viitataan usein suureen datamäärään, joka lisääntyy kiihtyvällä tahdilla ja on rakenteeltaan monimuotoista. Sillä tarkoitetaan yleensä valtavan kokoisia luokittelemattomia tietomassoja jotka voivat olla jopa tera- tai petatavun kokoisia (Kolehmainen 18.11.2011.) Tämänkaltainen data voi olla esimerkiksi tuotanto- ja toimitusprosessien tuottamaa automaattista seurantadataa, lokitietoa, sensorien lähettämää dataa (IoT) tai jotain muuta monimuotoista, ei rakenteellista dataa (Salo 2014, 35 - 36). Avoin data liitetään käsitteenä myös osaksi big dataa, tällöin puhutaan osin päällekkäin ja rinnan olevasta avoimesta datasta sekä myös linkitetystä datasta (Salo 2014, 43).

Yrityksillä on käytössään merkittävä määrä käyttökelpoista tietoa, joko omissa tai kumppaneiden tietovarastoissa, eri operatiivisissa järjestelmissä, pääte- ja mobiililaitteissa, intra- ja extranetissä, sähköisissä dokumenteissa, avoimena datana internetissä sekä sosiaalisessa mediassa (Niemelä 2018). Big Datan käsittelyä ja hyödyntämistä varten tämänkaltaisen datan käsittelyyn on kehitetty erilaisia algoritmeja hyödyntäviä tiedon louhintamenetelmiä ja koneoppimista, joiden avulla tästä tietomassasta voidaan tehdä erilaisia visuaalisia tilastollisia analyyseja (Vakkuri 20.6.2013).

Strukturoimattoman datan määrän eksponentiaalista kasvua kuvaa hyvin alla esitetty kuva, yksi eksatavu (EB) on 10^{18} tavua.



Kuva 10. Tiedon määrä ja kehitys eksatavuina (Holopainen 2016, 2)

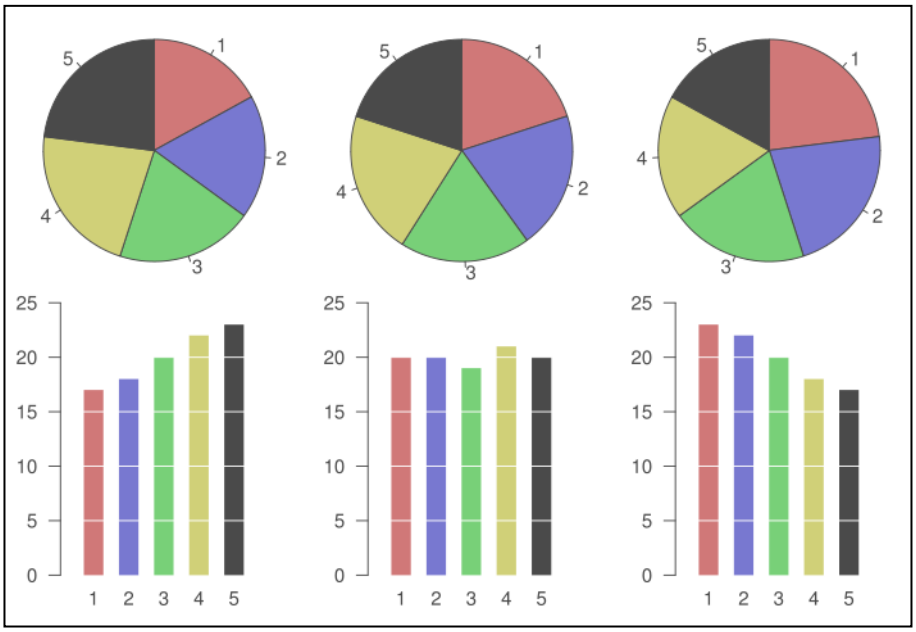
3.3 Data mart

Ilman spesifioituja datamartteja, moniulotteisia kuutioita tai metadata malleja, on tietovarastossa olevasta transaktio-, master-, tai raakadatasta hidasta ja tehotonta saada kustannustehokkaasti tarvittavia raportteja ja analyysejä. Datamartit voivat olla tähtimallin muotoon rakennettuja tietomalleja, kuutioita tai näppäriä leveitä tauluja, joko fyysisiä tai virtualisoituja (Hovi 17.5.2016). Eri liiketoimintajärjestelmistä ja joissain tapauksissa myös yrityksen ulkopuolelta kerättyä dataa puhdistetaan, aggregoidaan ja yhdenmukaistetaan keskistetystä tietovarastosta pienempiin tietomalleihin, Data Mart -rajapintoihin (Ilchenko 28.2.2017). Näitä datamartteja voidaan hyödyntää hyödyntämisrajapinnan työvälineillä, pilvipalvelumallisina tai lokaaleina itsenäiskäyttöön tarkoitetuilla raportointityökalulla (Self-Service reporting tool).

3.4 Visualisointimenetelmiä

Datavisualisointia toteutettaessa on syytä pohtia visualisoinnin sisältöä ja tavoitetta. Mitä visualisoinnilla halutaan tuoda esiin ja mille kohderyhmälle se on tarkoitettu. Visualisoinnilla voidaan mennä metsään jos yritetään ilmentää liikaa informaatiota yhteen ja samaan visualisointinäkymään. Tiedon välittämiseen visuaalisin keinoin on hyvä soveltaa graafisen suunnittelun ja hahmopsykologian oppeja. Seuraavan sivun kuvasta (Kuva 11, 15)

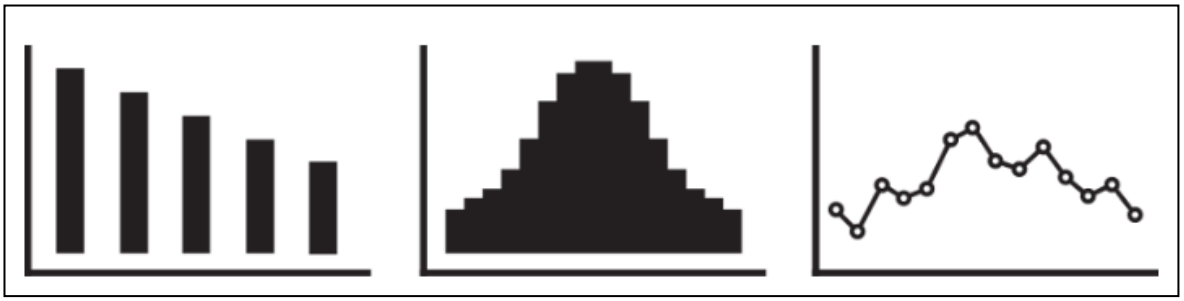
voimme todeta, että pylväskaavioiden esitystavalla on helpompi hahmottaa vertailua kuin ympyrädiagrammin avulla. Tämä pohjautuu aivojemme kykyyn hahmottaa pituuseroja helpommin kuin pinta-alaeroja (Koski 3.3.2015.)



Kuva 11. Pituuserojen hahmottaminen (Koski 3.3.2015)

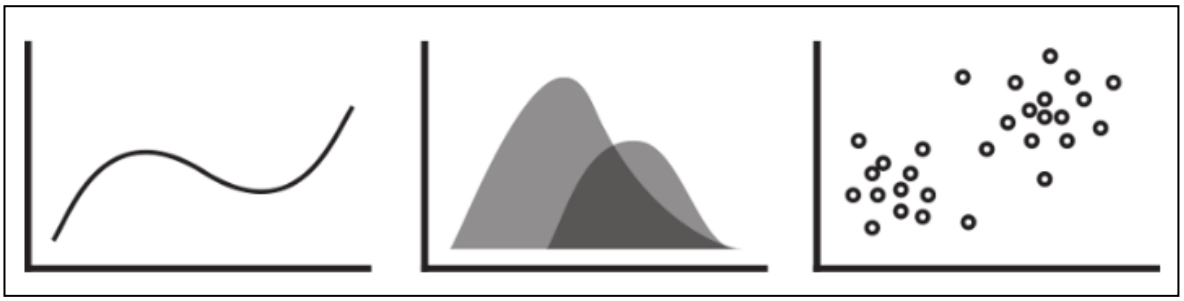
Esimerkiksi kokoomaraportteja tai analyysiesityksiä koostettaessa, on syytä huomioida graafisen suunnittelun perusteita eli sommittelua, typografiaa sekä värioppia (Koski 3.3.2015.)

Visualisoitavan datan sisältö ja määrä määrittelevät oikeaoppisen visualisoinnin esitystavan, kuten alla olevissa kuvissa 12 ja 13 on nähtävissä.



Kuva 12. Pylväskaavio (bar chart), Histogrammi (histogram), Viivakaavio (line chart) (Koski 3.3.2015)

Lisäksi alla olevat kuvat 13-15 ovat esimerkkejä tyypillisistä visualisointimalleista, joita yleisimmin käytetään datavisualisoinnissa.



Kuva 13. Funktion kuvaaja (function graph), Pinta-alakaavio (area chart), Pistekaavio (scatter plot) (Koski 3.3.2015)

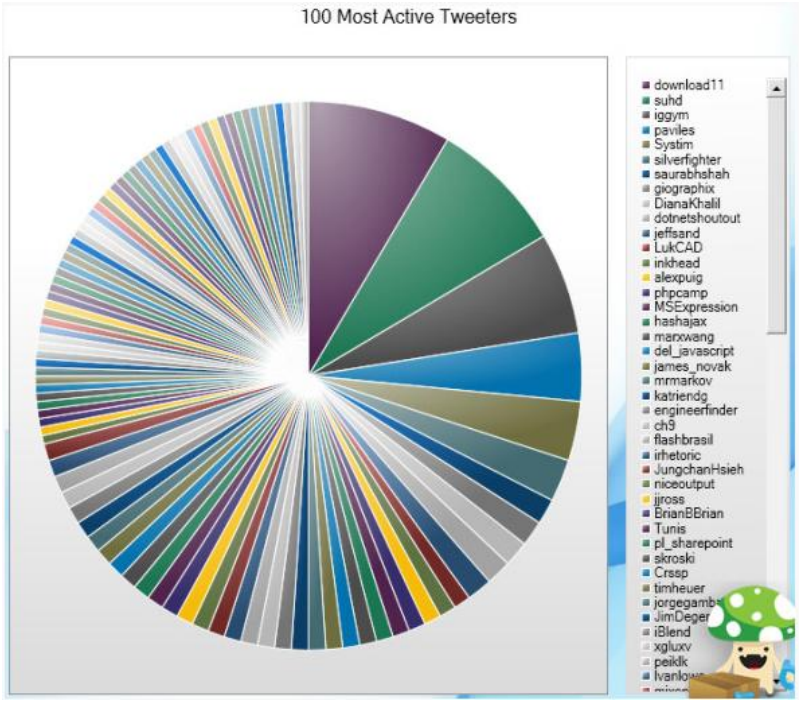
Kuva 14. Kuplakaavio (bubble chart), Kartogrammi (cartogram), Vuokaavio (flow chart) (Koski 3.3.2015)

Kuva 15. Venn-diagrammi (Venn diagram), Ympyrädiagrammi (pie chart), Tutkadiagrammi (Radar chart) (Koski 3.3.2015)

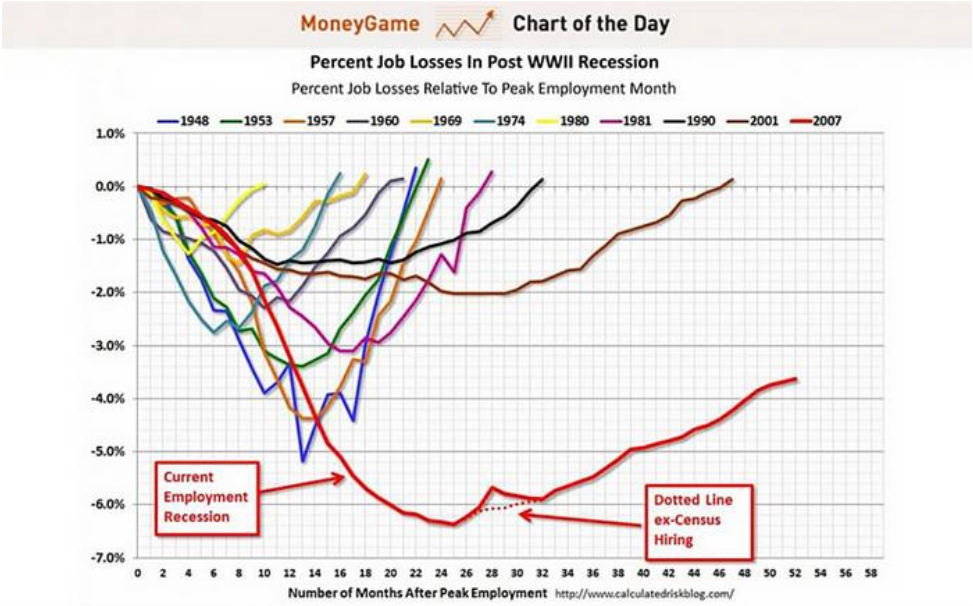
3.5 Visualisoinnin kompastuskivet

Tärkeä tekijä visualisointia suunniteltaessa on löytää oikeanlainen esitystapa tiedolle mitä halutaan viestiä eteenpäin. Mikäli yritetään sisällyttää liikaa tietoa yhteen kuvaan, sen viesti ja sanoma eivät nouse esille ja lopputuloksena on näkymä, jonka tulkinta ei ole nopeaa ja yksiselitteistä. Seuraavat kuvat 16 - 20 ovat esimerkkejä ns. huonoista data-visualisointitoteutuksista (Kuvat 16 - 20, 17 - 19).

Alla näkyvässä kuvassa numero 16 on visualisoinnissa käytetty ympyräkaaviota todella monen muuttujan kuvaamiseen. Kaavion viipaleen suhde kokonaisuuteen jää vaikeaksi hahmottaa sekä viipaleiden yksilöinti värikartasta on todella haastavaa.



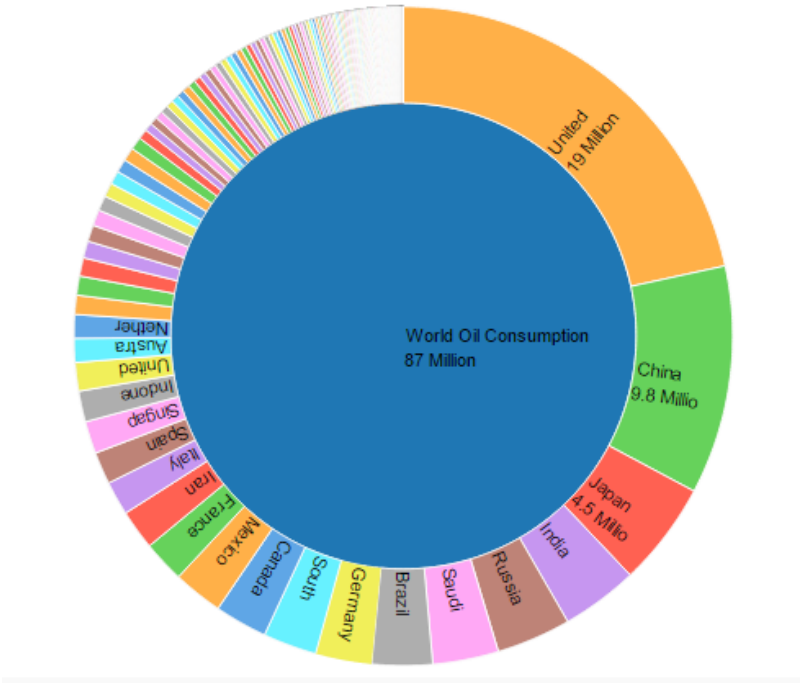
Kuva 16. Epäonnistunutta visualisointia, liian monen ulottuvuuden käyttö (IoTalents 2018)



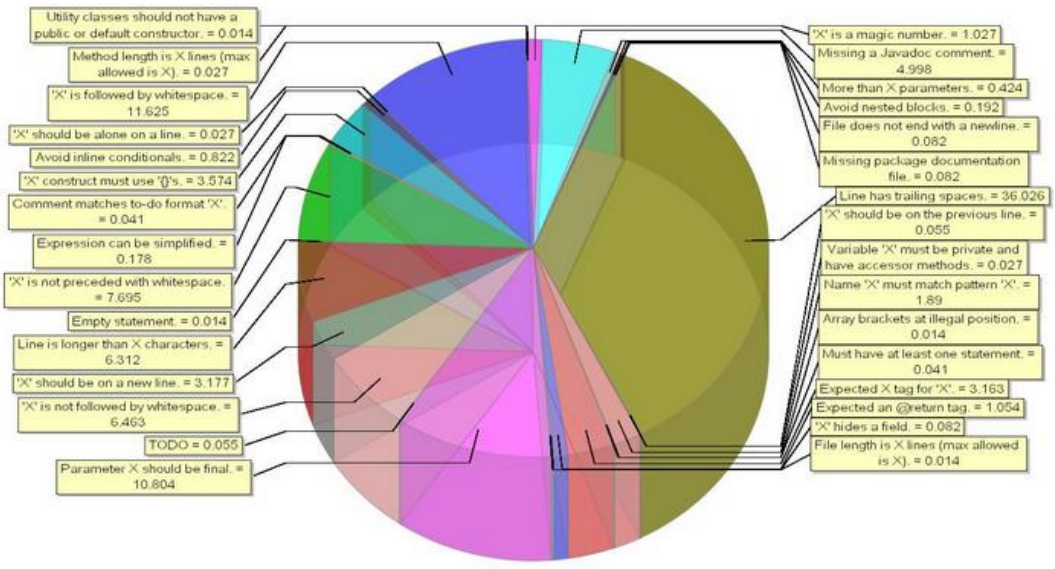
Kuva 17. Epäonnistunutta visualisointia, epäselvän sanoman esittäminen (IoTalents 2018)

Kun esitettävä asia on monimutkainen, niin visualisointikaan ei auta katsojaa ymmärtämään näkymän merkitystä (Kuva 17). Näkymässä on käytetty vaikeaselkoisia akselimuuttujia sekä liian monta värimuuttujaa.

Alla näkyvässä esimerkissä (kuva 18) on aivan liikaa muuttujia liitetty yhteen näkymään. Kaikkiin viipaleisiin ei edes mahdu muuttujan nimeä sekä näkyvät nimet ovat vaikeasti luettavissa.



Kuva 18. Epäonnistunutta visualisointia, asemointi ja väriongelmat (IoTalents 2018.)



Kuva 19. Epäonnistunutta visualisointia, ylikorostettu informatiivisuus (IoTalents 2018.)

Oheisessa esimerkissä (kuva 19) on yritetty kertoa liian paljon eksaktia informaatiota yhdessä ympyrädiagrammissa.

Tummasta taustasta on vaikea hahmottaa oleellinen viesti, kuten alla näkyvästä kuvasta 20 voi todeta. Tässä esimerkissä on myös vaikea erotella ympyrädiagrammin viipaleiden samankaltaisia värisävyjä.

Kuva 20. Epäonnistunutta visualisointia, värien käyttö (IoTalents 2018.)

3.5.1 Värien varomaton käyttö

Värejä käytettäessä on Sinkkosen ym. (2006, 132 - 133) mukaan syytä noudattaa seuraavia sääntöjä:

- Jos halutaan käyttäjän muistavan värien merkityksen, on värien maksimimäärä 5 + 2 väriä.
- Mikäli haluaa antaa väreillä määrämerkityksen, olisi hyvä käyttää spektrin järjestystä, joka on punainen, oranssi, keltainen, vihreä ja sininen.
- Väreihin liitetään myös syvyysvaikutus, jossa puhtaat, tummat tai lämpimät värit ovat lähimpinä, eivätkä näin ollen sovellu taustaväreiksi.
- Kirkkaat värit mielletään huomiota herättäviin, muistuttaviin ja vaarasignaaleihin.
- Vierekkäin ei kannata sijoittaa spektrin ääripäiden värejä, kuten sinistä ja punaista.

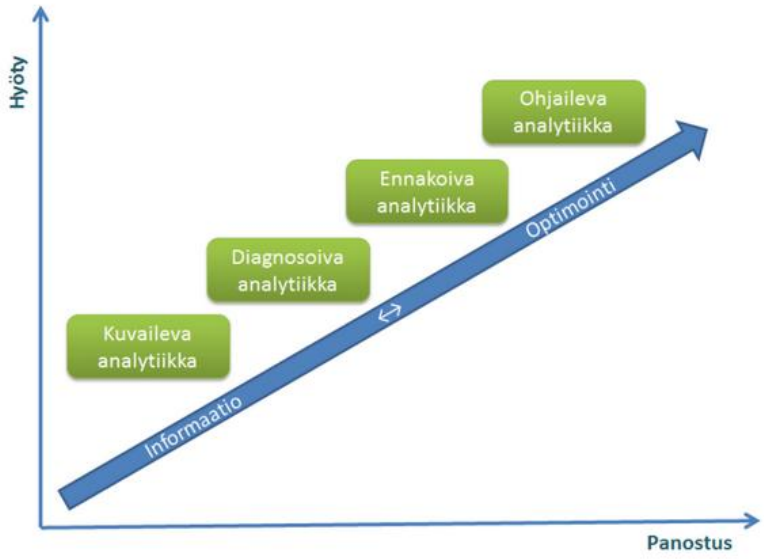
Värisokeiden huomioimiseksi Sinkkonen ym. (2006, 134) suosittelee ottamaan huomioon seuraavat suunnitteluohjeet:

- Lisää monivärisiin kuvakkeisiin tekstivihjelaatikko tai selite.
- Vältä käyttämästä punaista, vihreää, ruskeaa, harmaata sekä sinipunaista väriä vierekkäin tai keskenään muuttuvina väreinä.
- Muista kontrastien käyttö, pois lukien tausta ja teksti, myös kuvien sisällä.
- Värisokean on helpointa erottaa keskenään sininen, keltainen, musta ja valkoinen.
- Vältä käyttämästä värisignaaleja, kuten värien vaihtumista punaisesta vihreäksi, punaisesta keltaiseksi tai vihreästä keltaiseksi (liikennevalot).

4 Datavisualisointi ja analytiikan hyödyntäminen

Datan visualisointiprosessi nähdään usein osana laajempialaista analysointia, jossa dataa tutkitaan ensin visualisointiohjelmistolla, pyritään löytämään relaatioita muuttujien välillä ja lopuksi koostetaan yhtenäinen näkymä datasta. Kun visualisointia hyödynnetään data-analyysiprosessin tukena, voidaan datasta havaita huomattavasti helpommin relaatiot sekä korrelaatiot muuttujien välillä, joiden varaan voidaan alkaa soveltamaan edistyneempää analytiikkaa, data-analyysin jatkoksi, toteaa Väisänen (15.8.2018.)

Vuonna 2004 alkanut trendi interaktiivisuuteen ja visuaalisuuteen perustuvien ketterän analytiikan itsepalveluperusteisien BI-ratkaisujen suosiolle ei näytä laantuvan, sillä Gartnerin (Liite 1, 1) mukaan vuoteen 2020 mennessä markkinoita tulevat hallitsemaan BI sektorilla automaattiseen tiedonjalostukseen perustuvat visuaaliset tiedon analyysityökalut, joissa hyödynnettäisiin koneoppimista sekä edistynyttä analytiikkaa.



Kuva 21. Analytiikan tasot (Datatiede 2014)

Data-analytiikalla tavoiteltavaa hyötyä voidaan tarkastella kuvan 21 mukaisesti analytiikan sovellusten skaalautuvuudella. Eri tasoja hyödyntäen on yrityksessä tai organisaatiossa mahdollista tarkastella miten data-analytiikkaa hyödynnetään nyt ja tulevaisuudessa (Datatiede 2014.) Analytiikan tasojen sisältöä kuvataan Datatiede (2014) sivuston mukaan seuraavasti:

- Kuvaileva analytiikka sisältää tilastotiedettä, tutkivaa data-analyysiä ja visualisointia. Sen päämääränä on selvittää tapahtumien syitä sekä arvioida tuloksia.
- Diagnosoiva analytiikka keskittyy tarkastelemaan olosuhteita jotka johtivat em. tapahtumiin, eli vastaamaan kysymykseen ”miksi näin tapahtui?”. Siinä analysoidaan

aikaisempia tuloksia syy-seuraussuhteiden havaitsemiseksi syvällisen analysoinnin kautta.

- Ennakoiva analytiikka keskittyy tulevien tapahtumien skenaarioiden kautta ennakoimaan tapahtumia. Siinä esitetään ennusteita, luokitteluja ja assosiaatioita erilaisten tulevaisuuteen perustuvien tapahtumamallien kautta.
- Ohjailevan analytiikan avulla etsitään vastauksia kysymykseen ”mitä nyt pitäisi tehdä?”. Siinä pyritään löytämään parhaat toimintamallit erilaisten päätöspuuteorioiden, simulaatioiden ja optimoinnin avulla sekä hyödyntää niistä saadut tulokset automaattisesti.

4.1 Gartnerin nelikenttä BI työkaluista

BI analytiikan markkinoita ja alan toimijoita seuraava on oletettavasti tietoinen alla näkyvästä Gartnerin nelikentästä (kuva 22). Gartner on Yhdysvaltalainen markkinatutkimus- ja konsulttiyritys jonka vuosittain julkaisemilla Business Intelligence -alan markkinatilanteen nelikenttäanalyyseillä (Magic Quadrant) on paljon painoarvoa yritysten investoidessa BI teknologiahankintoihin. Nelikentässä eri alan toimijoita sijoitellaan 13 erilaisen välineen käyttökokemuksien ja kehitysaktiivisuuteen perustuvien kriteerien perusteella (Pengon 28.10.2015.)

Kuva 22. Gartnerin nelikenttäanalyysi BI ja analytiikkatoimijoista vuonna 2018 (LIITE 1)

Nelikentän arviointikriteereitä on kaksi, pystyakselilla ”Ability to execute”, joka tarkoittaa ohjelmiston suorituskykyä, käyttäjäystävällisyyttä sekä asiakastyytyväisyyttä sekä vaaka-akselilla ”Completeness to vision”, joka tarkastelee toimijan monipuolisuutta, eli löytyykö

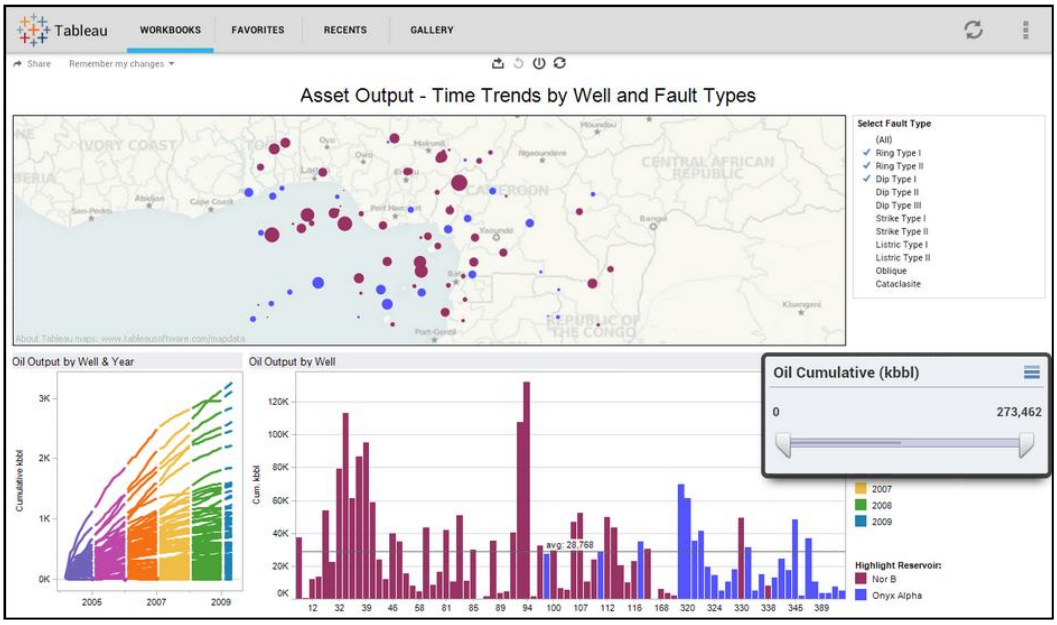
tuoteperheestä tuotetta eri käyttöalustoille sekä esimerkiksi edistyneeseen analytiikkaan ja budjetointiin (Niemijärvi 2.4.2013.) Arviointikriteerien perusteella ohjelmistot luokitellaan Niemijärven mukaan (2.4.2013) neljään eri ryhmään:

- Niche players (pienen markkinan toimijat)
- Visionaries (visionääreihin)
- Challengers (haastajiin)
- Leaders (johtajiin).

4.2 Tableau

Tableau (Kuva 23, 22) on noussut nopeasti kärkeen BI työkalujen markkinoilla. Se on Yhdysvaltalainen ohjelmistoyhtiö, joka on perustettu vuonna 2003 ja sen toiminta keskittyy datavisualisointeihin ja liiketoimintatiedon hallintaan. Yritys tutki alkutaipaleellaan datavisualisointitekniikoita Yhdysvallan puolustusministeriölle ja kaupallistivat toimintansa kyseisen tutkimuksen pohjalta.

Tableau Desktop interaktiivista datavisualisointityökalua markkinoidaan käyttäjäystävällisenä tuotteena, jonka käyttämiseksi ei tarvitse osata ohjelmointikieliä. Se tarjoaa valmiit datayhteydet teksti- ja Excel- ja statistiikkatiedostoista aina suoriin tietokantayhteyksiin sekä big data alustoihin erittäin kattavasti. Tarvittaessa voi käyttää myös ODBC-rajapintaa yhteyden luomiseen. Lähdedatan sisäänluvun jälkeen voi dataa muokata, yhdistellä sekä suodattaa ennen visualisoinnin kehittämiseen siirtymistä. Datavisualisointien teko tapahtuu periaatteessa ”drag and drop” tyylisesti ohjelmiston pyrkiessä taustalla automaattisesti muodostamaan parhaan mahdollisen oletusvisualisoinnin käyttäjän valitsemasta datasta.

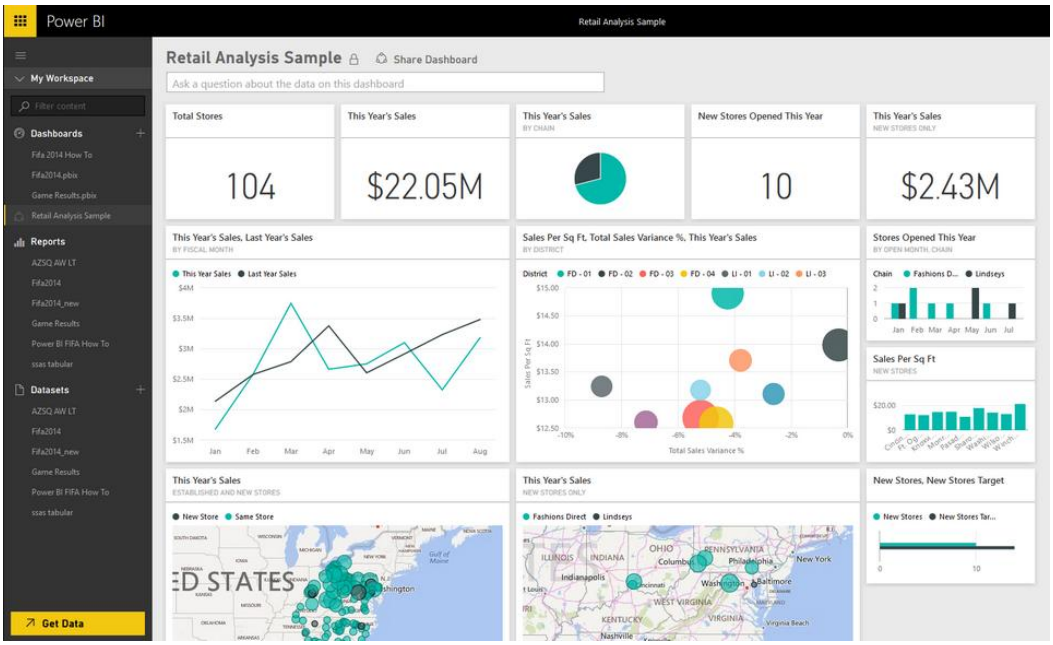


Kuva 23. Tableau ohjelmistolla toteutettu ”dashboard” (Aerow 13.4.2017)

Tableaussa on myös sisäänrakennettu karttapohja, jota voidaan hyödyntää paikkatietoon perustuvan visualisoinnin yhteydessä (Kuva 23, 22). Se tunnistaa sisäänluettavasta datasta koordinaattitiedon, kaupungin tai paikan nimiä sekä postinumerot. Tableau visualisointeja voi koostaa interaktiivisiin dashboardeihin sekä story -pohjaisiin esityksiin, jotka muistuttavat hieman perinteistä MS Powerpoint -esitystä.

4.3 Power BI

Power BI on kuvattu seuraavassa kuvassa 24, joka on Microsoftin vuonna 2015 julkaisema tiedon visualisointiin perustuva raportointi- ja analysointityökalu. Siinä on myös monipuoliset mahdollisuudet tuoda dataa eri lähteistä ja yhdistellä niistä tietomalleja DAX-ohjelmointikieltä hyödyntäen. Se on lyhyessä ajassa noussut vahvaksi tekijäksi alan markkinoilla, johtuen käyttäjäystävällisyydestä sekä edullisesta hinnoittelusta ja sen käyttö onkin yleistynyt analyytikkojen ja kontrollereiden keskuudessa perinteisen Excelin rinnalla (Enho 30.1.2016.)

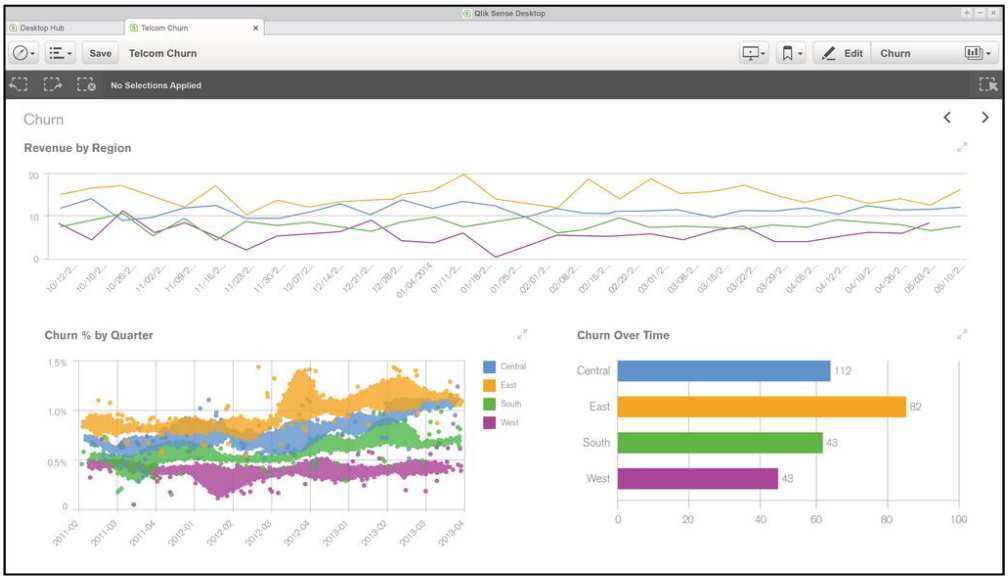


Kuva 24. Power BI ohjelmistolla toteutettu ”dashboard” (Aerow 13.4.2017)

4.4 QlikSense

Ohjelmistoyhtiö Qlik julkaisi vuonna 2014 itsepalveluraportointityökalun QlikSensen (Kuva 25, 24). Siinä on hyvin paljon samankaltaisuutta kuin kilpailija Tableaussa. Se on nimenomaan datan visualisointiin tarkoitettu helppokäyttöinen ohjelmisto, jolla voi luoda oma-

aloitteisesti raportteja, visualisointeja sekä analyysinäkymiä. Siihen voi myös yhdistää dataa eri lähteistä ja analysoida sekä yhdistää ristiin eri lähteiden tietoja sekä käyttää myös mobiililaitteilla (Qlik 18.9.2014.) Qlik yhtiö on perustettu v.1993 Ruotsissa ja se on saavuttanut vankan markkina-aseman pohjoismaiden BI raportoinnin ja analytiikan sektorilla, ”raskaamman” raportointialustansa, QlikView:n myötä. QlikSensen käyttäjät ovat kehuneet sen ”drill-down” eli hierarkisen tiedon porautumis-ominaisuuden helppoutta.



Kuva 25. QlikSense ohjelmistolla toteutettu dashboad (Aerow 13.4.2017)

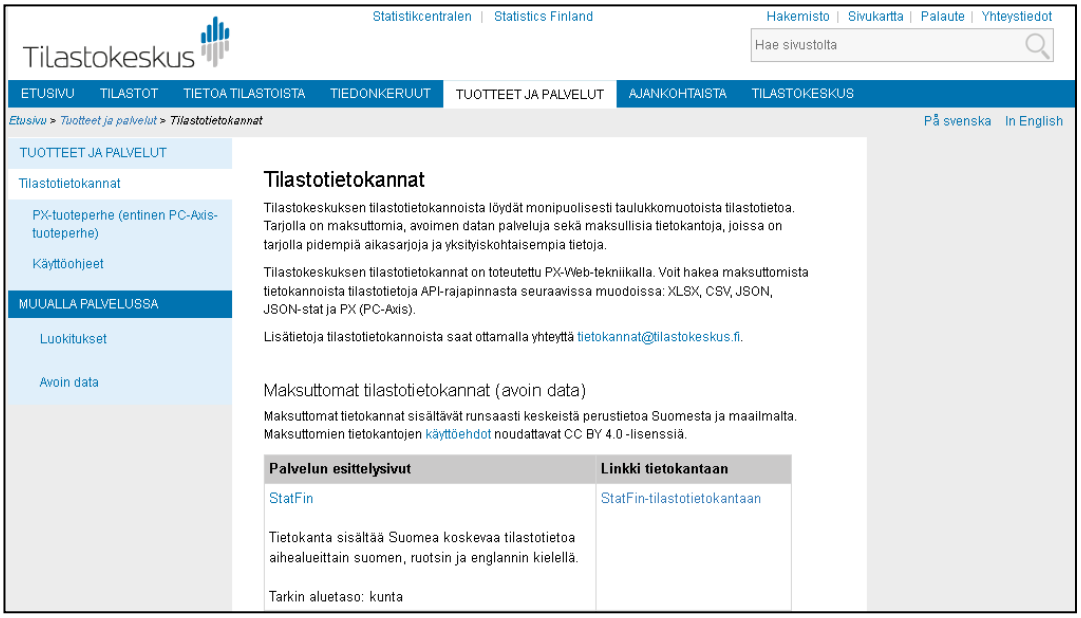
Yhteistä näille yllä esitetyille datavisualisointiohjelmistoille on mahdollisuus liittää niihin R, avoimen lähdekoodin ohjelmointiympäristö, jolla voidaan toteuttaa edistyneempää analytiikkaa. R on maailmanlaajuisesti tunnettu tilastollisille malleille suunniteltu ohjelmointikieli, jota hyödynnetään myös ennusteanalyysien tuottamisessa suuristakin datamassoista (Storås 18.2.2015). R ohjelman kautta saadut laskentatulokset voidaan tuoda esitettäviksi visualisointiohjelmille ja liittää näkymiin.

5 Case: Pilvipalveluiden hyödyntäminen yrityksissä

Toimeksiantona tässä case-esimerkissä on visuaalisen näkymän toteuttaminen Tableau ohjelmistolla kotimaisten yritysten, suuruusluokkaa yli 100 henkilöä/yritys, pilvipalveluiden käyttöönoton laajuudesta 5 - 6 vuoden aikajaksolla. Lähteenä käytetään avointa dataa Tilastokeskuksen PX-Web-tietokannasta.

5.1 Lähdeaineiston hankinta

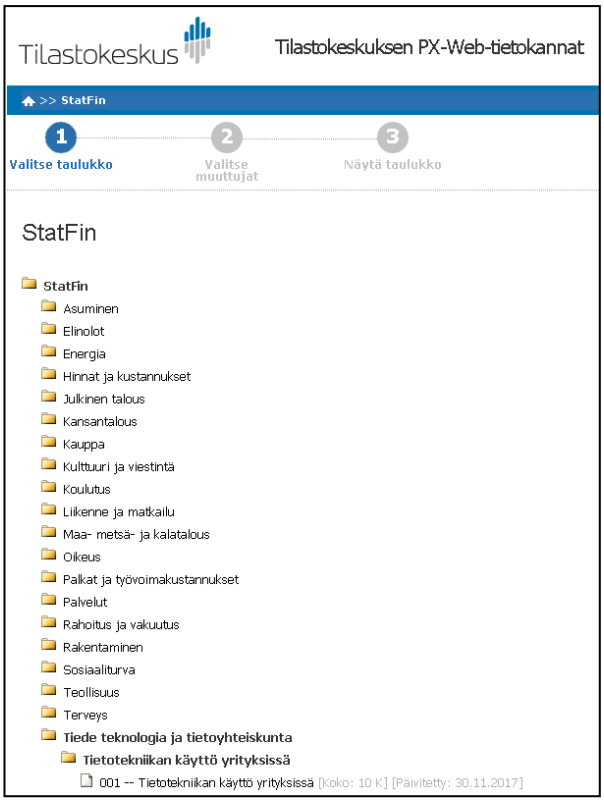
Tilastokeskuksen julkaisemaa avointa dataa voi ladata vapaasti esimerkiksi heidän tarjoamaltaan StatFin-tilastotietokannan alustalta. Tämä tietokanta sisältää eri aihealueittain Suomea koskevaa tilastotietoa, kuten alla näkyvästä kuvasta voidaan todeta.



Kuva 26. Tilastokeskuksen tilastotietokannat, tuotteet ja palvelut (Tilastokeskus 2018)

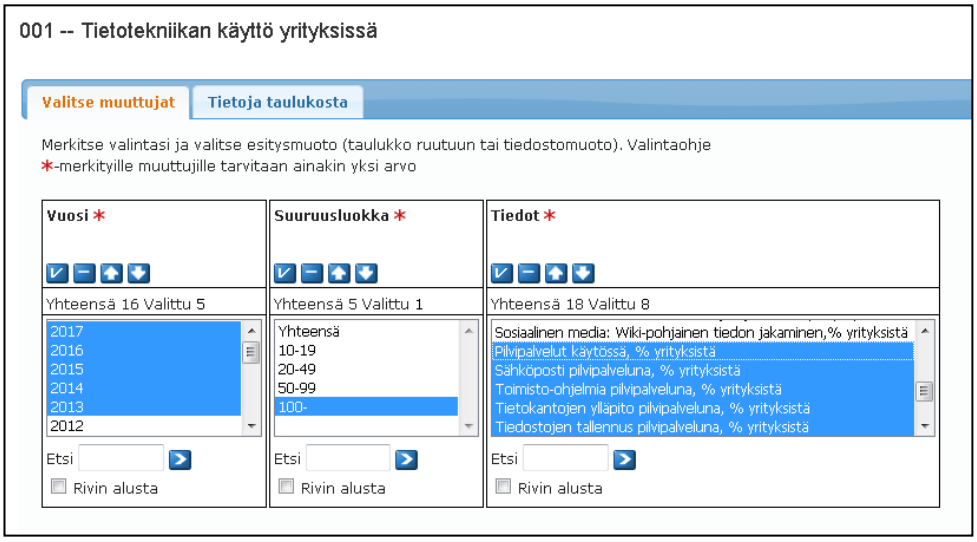
Tilastokeskus tuottaa avoimen datan käyttäjille valmiita rajapintoja. Opinnäytetyötä kirjoittaessa rajapintojen luominen oli aloitettu StatFin-tietokannasta ja sitä oltiin laajentamassa vaiheittain. Tässä case-esimerkissä käytetty data on uudesta PX-Web tietokannasta joka käyttää PX-Web API rajapintaa. Se mahdollistaa avoimen tilastodatan hakemisen koneellisesti xlsx, csv, json, json-stat sekä px(PC-Axis) -formaateissa. Vastaavanlainen toteutus on käytössä myös Ruotsin tilastovirastossa (Tilastokeskus 2018.)

Pilvipalveluiden käyttöä kuvaava aineisto löytyy tilastosta ”Tietotekniikan käyttö yrityksissä” (Kuva 27). Tilastossa kuvataan suhdelukuina yritysten tietotekniikan ja sähköisen liiketoiminnan käyttöastetta.



Kuva 27. Tilastokeskuksen PX-Web-tietokannat, StatFin taulukot (Tilastokeskus 2018)

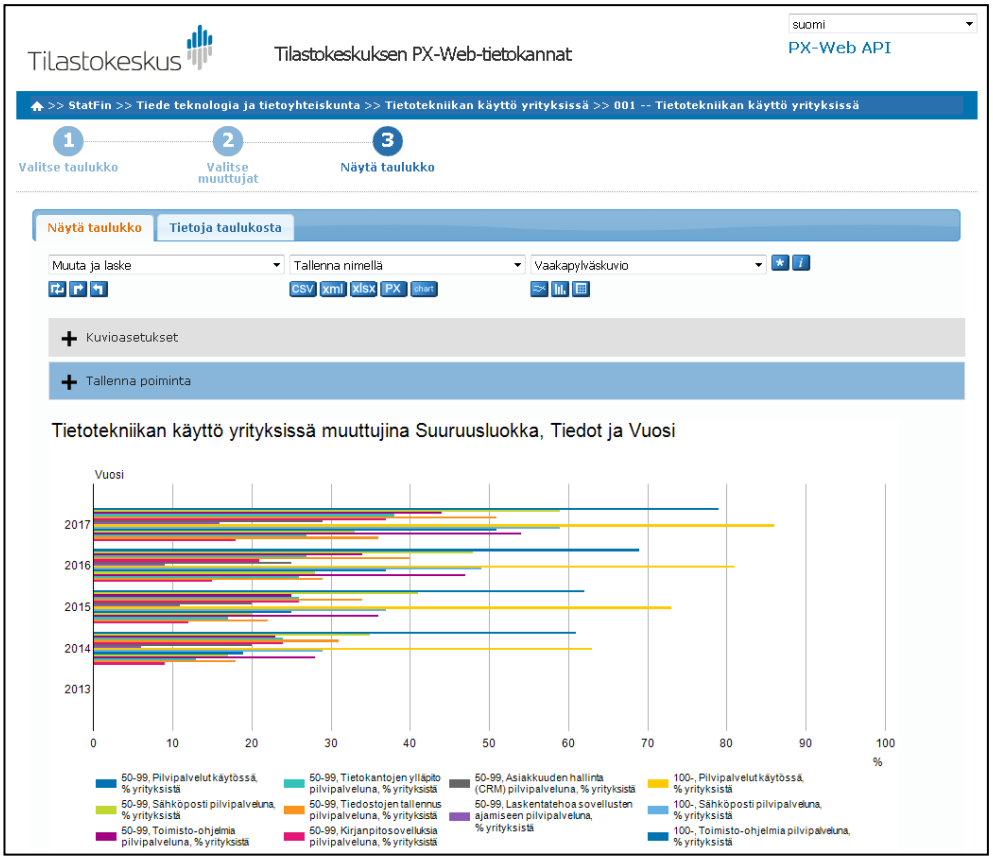
Aineistosta kuvan 28 mukaisesti valitaan vuosi, suuruusluokat sekä tarkasteltavat tiedot, tässä tapauksessa pilvipalveluiden käyttöä kuvaavat attribuutit.



Kuva 28. Tilastokeskuksen PX-Web-tietokannat, muuttujien määrittely (Tilastokeskus 2018)

Koska vuodelta 2012 ei tilastoitua dataa ole pilvipalveluiden osalta saatavilla (Kuva 28, 26), rajataan otanta koskemaan vain vuosia 2013 - 2017.

Tiedosto kannattaa ladata csv -formaatissa, koska tällöin sisältö saadaan ryhmittelemättömänä ja on näin ollen valmiimpi sisään luettavaksi visualisointiohjelmistoon. Mikäli tiedosto halutaan xlsx -formaattina, joutuu sitä hieman muokkaamaan, poistaen ryhmittelyjä esimerkiksi vuosikentän osalta. Tilastokeskuksen PX-Web palvelu tarjoaa myös mahdollisuuden tarkastella aineistoa visuaalisesti jo suoraan heidän verkkosivullaan, mutta kuten alla näkyvästä kuvasta 29 nähdään, ei esitystapa ole kovin selkeä tälle aineistolle.



Kuva 29. Tilastokeskuksen PX-Web-tietokannat, aineiston lataaminen (Tilastokeskus 2018.)

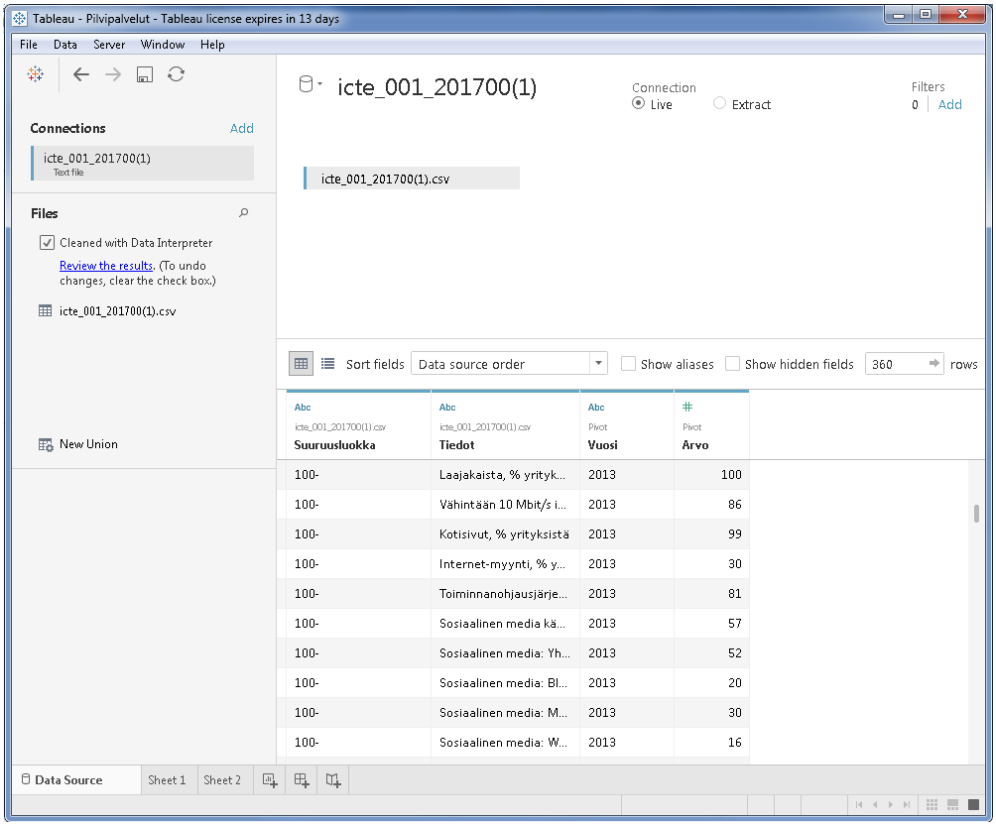
Ladataan valittu aineisto csv -tiedostona omalle työasemalle ja luetaan sisään tekstitiedostona Tableau -ohjelmistolla aineiston visualisointia varten.

5.2 Aineiston visualisointi Tableau ohjelmistolla

Sisään luetulle tiedostolle voidaan tehdä muutoksia ”Data Source” välilehdellä (kuva 30). Kenttiä voidaan pivotoida sekä muuttaa nimiä ja tietotyyppiä. Myös useampia lähdetiedos-

toja voidaan käyttää samanaikaisesti, jolloin tiedostoille tarvitsee antaa liitosavaimet sekä liitosten tyyppimäärittely. Tässä esimerkissä on hyödynnetty Tableaun sisäänrakennettua automaattista datatulkkia tiedon puhdistamiseen (Data Interpreter), se poistaa automaattisesti esimerkiksi ylimääräiset tyhjät- sekä seliterivit, joita avoimen datan tiedostoissa tulee usein mukana.

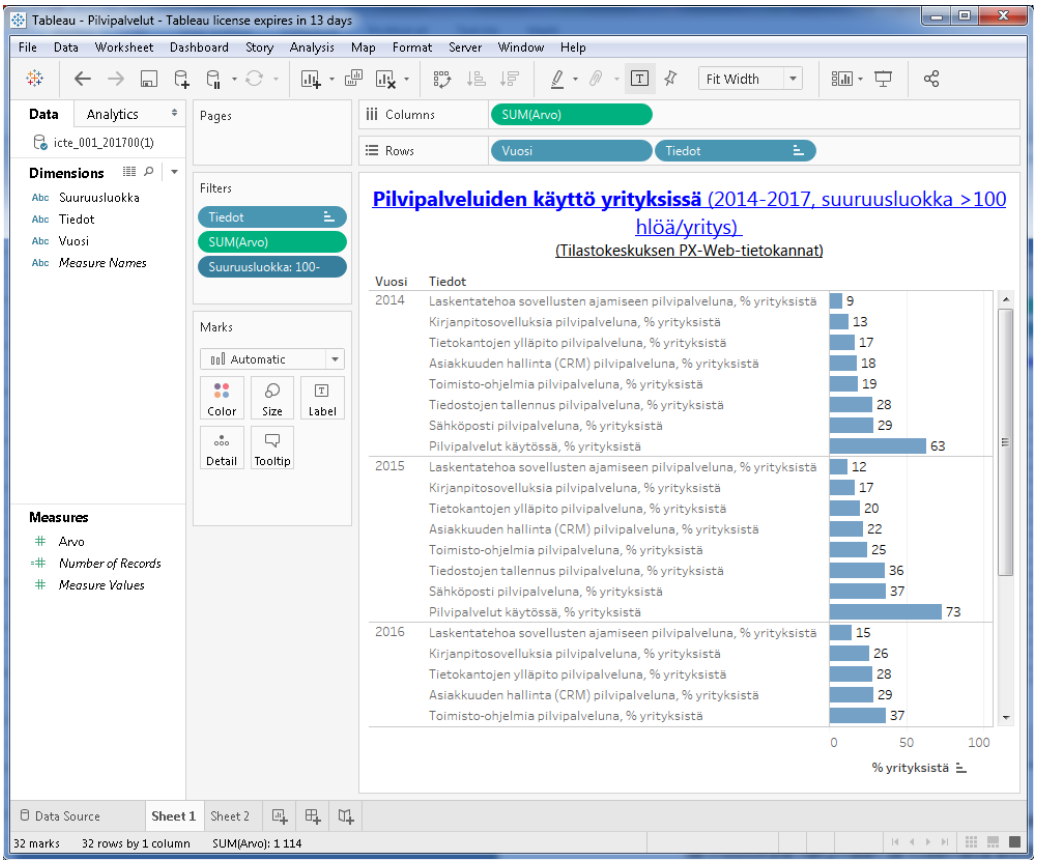
Alla näkyvän kuvan yläreunassa oleva yhteydenmäärittely (Connection) käyttää suoraa yhteyttä lähdetiedostoon. Mikäli kyseessä olisi tietokantayhteys, niin suositellumpi vaihtoehto olisi käyttää ”Extract” ominaisuutta, joilloin ohjelmisto lataa tietokannasta kompressoidun kopion työaseman muistiin snapshot -tyyppisenä TDE -tiedostona (Tableau Data Extract). Suora yhteys usein hidastaa ohjelmistoa ja tiedoston päivityskyselyt ovat riippuvaisia tietokannan kuormituksesta ja suorituskyvystä.



Kuva 30. Lähdeaineiston lataus Tableau desktop ohjelmistoon.

Kun datan muokkaus on saatu tehtyä, voidaan siirtyä tekemään visualisointia ”Worksheet” välilehdelle kuvan 31 mukaisesti (Kuva 31, 29). Näitä välilehtiä voi rajattomasti lisätä ja tehdä niille erilaisia datavisualisointeja, joita voidaan sitten myöhemmin liittää koostesivuille (Dashboard) tai esityssivuille (Storyboard).

Huomioi seuraavasta kuvasta 31 ”Marks” paneelin muotoiluvaihtoehdot (Color, Size, Label, Detail ja Tooltip), joihin voidaan raahata tieto -objekteja vasemman reunan Data -valikosta. Värimuotoilua (Color) tulen myöhemmin hyödyntämään tässä toimeksiannossa. Luodun työnäkymän (Worksheet) voi myös kopioida uuden näkymän pohjaksi ja tehdä siihen lisäyksiä tai muutoksia tarvitsematta aloittaa koko näkymän tekoa alusta. Tämä on myös hyödyllinen tapa säästää aikaa, kun halua tehdä erilaisia kokeiluja ja vaihtoehtoja löytääkseen optimaalisimman esitystavan näkymälleen.

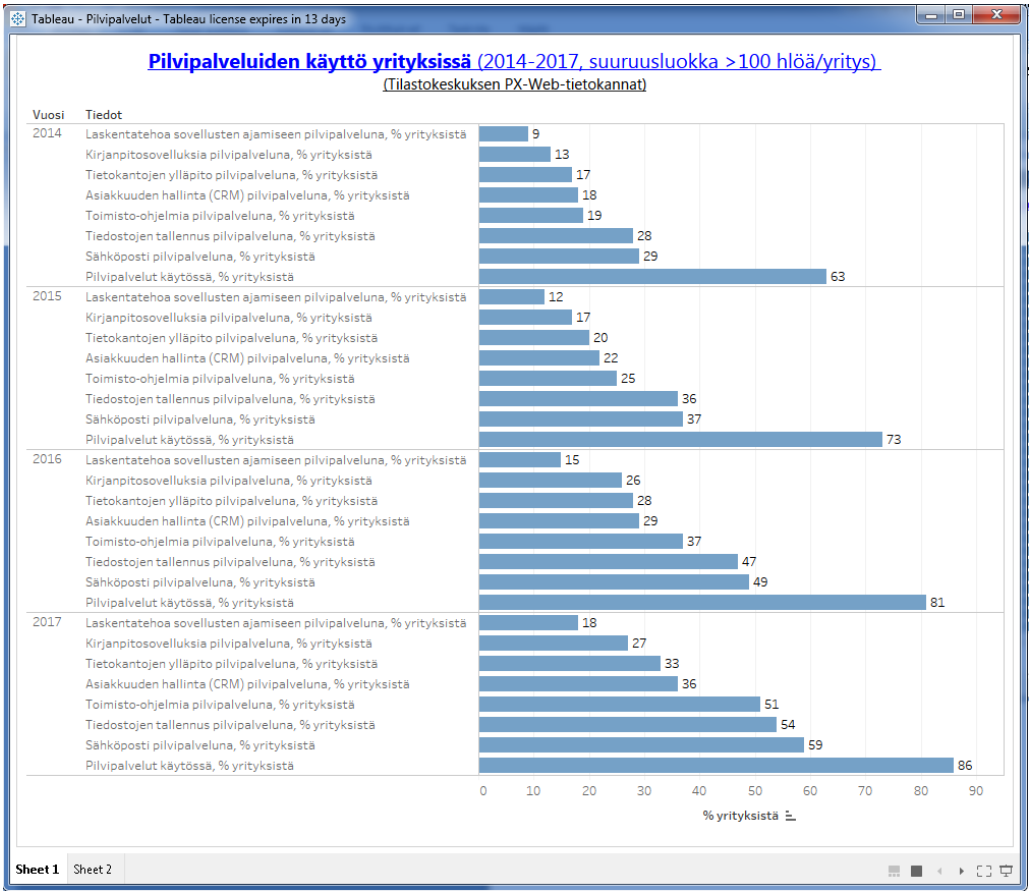


Kuva 31. Lähdeaineiston visualisointi.

5.3 Tulokset

Seuraavilla sivuilla esitetyt kuvat 32 – 34 esittävät erilaisin visuaalisin esitystavoin toimeksiannon tuloksena syntyneet näkymät. Kuvista voimme havaita, että pilvipalveluiden käyttö yrityksissä, joiden kokoluokka on yli 100 henkilöä, on kehittynyt tasaisesti vuosittain. Pilvipalveluita on käytössä jollain osa-alueella jopa 86 % yrityksistä. Vuonna 2017 ylittyi jo 50 %:n raja sähköpostipalveluiden, tiedostojen tallennuksen- sekä toimisto-ohjelmien käytössä pilvipalveluina.

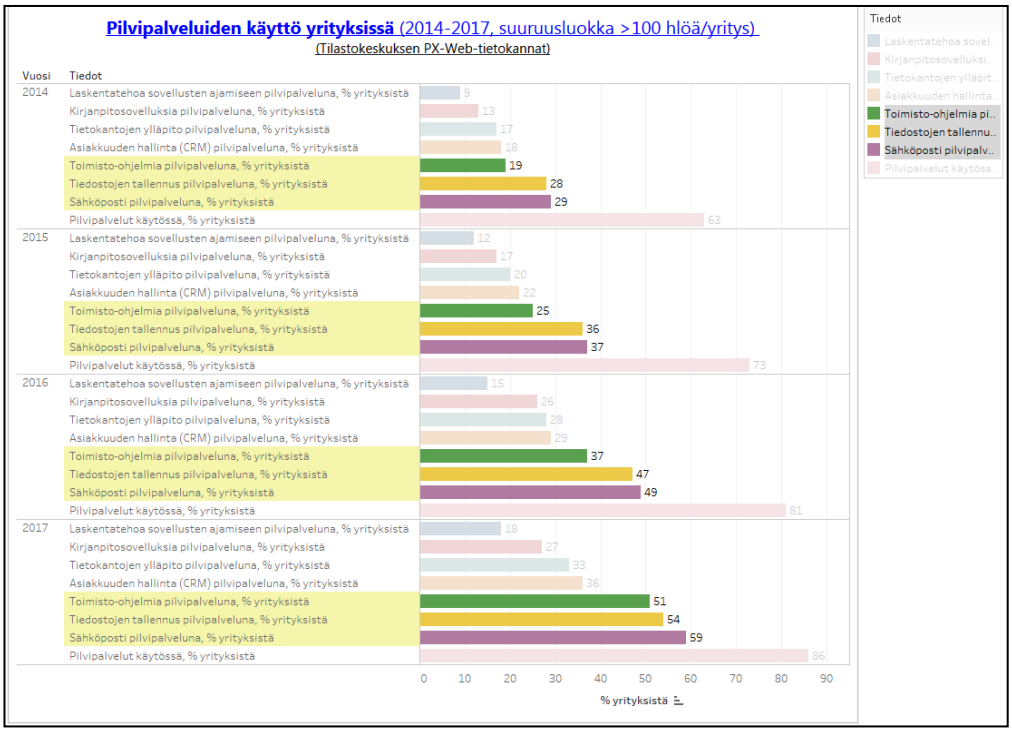
Vähiten hyödynnetään laskentatehoa sovellusten ajamiseen pilvipalveluna, vain 18 % vuonna 2017, tosin sen määrä on kaksinkertaistunut vuodesta 2014, joka on nähtävissä alla olevasta kuvasta.



Kuva 32. Pilvipalveluiden käyttö yrityksissä (v. 2014 - 2017), yli 100 hlöä/yritys.

Kuvan 32 näkymässä ovat aineiston tiedot sijoitettuna vaakasuuntaisen pylväsdiagrammin muotoon, jossa tiedot on ryhmitelty vuositasolla ja prosentuaaliset käyttäjämäärät eri tietokategorioittain. Yhtä värimallia käytettäessä nähdään suuruuserot helposti, ja kun näkymään on vielä lisätty myös prosenttiarvon lukema, niin sitä ei tarvitse tulkita erikseen akselin arvoasteikolta. Vuosittainen kasvu on nopeasti havaittavissa kaikilla osa-aluilla.

Alla näkyvä visualisointi (Kuva 33, 31) on rakenteeltaan vastaava kuin edellinen näkymä, mutta tässä on sijoitettu ”Tiedot” dimensio myös ”Marks” paneelin värimuotoilun alle, jolloin sisältöä voidaan korostaa oikeassa yläreunassa näkyvän tiedot/väri suodattimen avulla. Sen avulla voimme selkeämmin tarkastella kolmea suosituinta pilvipalvelutoimintoa ja havaita että vuosien 2016 -2017 välillä on eniten lisääntynyt edellisvuodesta toimisto-ohjelmien käyttö pilvipalveluna, peräti 14 prosentin kasvu edellisvuoteen. Sähköpostin käyttö yrityksissä pilvipalveluna jatkaa sen sijaan noin 10 prosentin kasvuvauhtia vuositasolla.



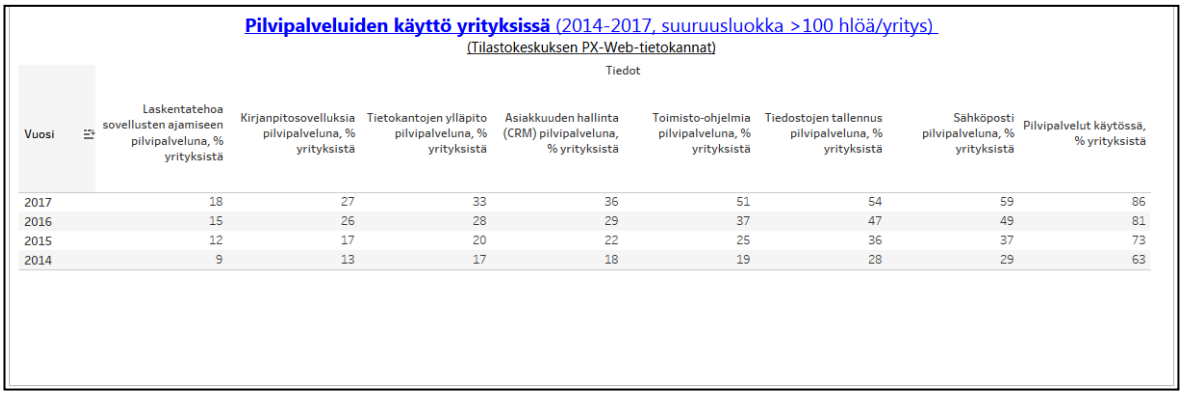
Kuva 33. Tieto -attribuutti väripaletin avulla esitettynä ja osittain korostettuna.

Kuva 34. Vuosi -attribuutin korostaminen väripaletin avulla.

Kun aikamuuttujaa käytetään värimuotoilun kautta, voidaan pylväsnäkymästä havaita eri vuodet ja prosenttiluvut omina väriosioinaan, kuten kuvasta 34 on havaittavissa. Väri/vuosi suodattimella (oik. yläreuna) voidaan myös haluttaessa korostaa näkymästä eri vuosien palkit, eli suodattimet toimivat interaktiivisesti. Kuvasta voidaan havaita Pilvipalveluiden kokonaisuudessaan hieman taantuneen vuositasolla 2014 – 2017, kumulatiivisen muutoksen ollessa +10, +8 ja +5 prosenttia vuosille 2015 - 2017.

Datavisualisoinnin avulla aineiston tulkinta nopeutuu moninkertaisesti. Seuraavassa kuvassa 35 on sama aineisto esitettynä perinteisenä taulukkonäkymänä. Kuten huomaam-

me, on analyysien tekeminen huomattavasti hitaampaa, kun visuaaliset efektit jätetään pois näkymästä. Katsojan silmä hakee lukuarvoista poikkeavuuksia ja aivoissa tapahtuu muistiprosessointia ja lukuvertailuja.

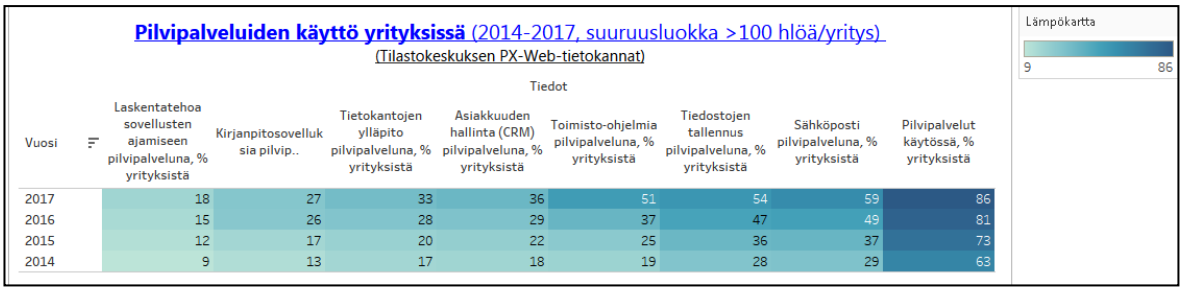


Pilvipalveluiden käyttö yrityksissä (2014-2017, suuruusluokka >100 hlöä/yritys)
(Tilastokeskuksen PX-Web-tietokannat)

Tiedot

Vuosi	Laskentatehoa sovellusten ajamiseen pilvipalveluna, % yrityksistä	Kirjanpitosovelluksia pilvipalveluna, % yrityksistä	Tietokantojen ylläpito pilvipalveluna, % yrityksistä	Asiakkuuden hallinta (CRM) pilvipalveluna, % yrityksistä	Toimisto-ohjelmia pilvipalveluna, % yrityksistä	Tiedostojen tallennus pilvipalveluna, % yrityksistä	Sähköposti pilvipalveluna, % yrityksistä	Pilvipalvelut käytössä, % yrityksistä
2017	18	27	33	36	51	54	59	86
2016	15	26	28	29	37	47	49	81
2015	12	17	20	22	25	36	37	73
2014	9	13	17	18	19	28	29	63

Kuva 35. Aineisto esitettynä perinteisenä taulukkonäkymänä.



Pilvipalveluiden käyttö yrityksissä (2014-2017, suuruusluokka >100 hlöä/yritys)
(Tilastokeskuksen PX-Web-tietokannat)

Tiedot

Vuosi	Laskentatehoa sovellusten ajamiseen pilvipalveluna, % yrityksistä	Kirjanpitosovelluksia pilvip..	Tietokantojen ylläpito pilvipalveluna, % yrityksistä	Asiakkuuden hallinta (CRM) pilvipalveluna, % yrityksistä	Toimisto-ohjelmia pilvipalveluna, % yrityksistä	Tiedostojen tallennus pilvipalveluna, % yrityksistä	Sähköposti pilvipalveluna, % yrityksistä	Pilvipalvelut käytössä, % yrityksistä
2017	18	27	33	36	51	54	59	86
2016	15	26	28	29	37	47	49	81
2015	12	17	20	22	25	36	37	73
2014	9	13	17	18	19	28	29	63

Lämpökartta
9 86

Kuva 36. Taulukkonäkymään lisättynä lämpökartta efekti.

Edellisestä taulukkonäkymästä saa välittömästi erottumaan oleellisen tiedon käyttämällä lämpökarttamallia tiedon visualisointiin, kuten yllä näkyvästä kuvasta 36 voimme havaita. Katsojan näköaistin välittämä tieto aivoille värien muutoksista näkymässä riittää nopean analyysin tekemiseen siitä mitä data haluaa meille kertoa. Lämpökarttaa käytettäessä on hyvä huomioida, että väripaletista käytetään vain yhden perusvärin sävyjä.

6 Pohdinta

Data visualisoinnin hyötynäkökulmia on helppoa ymmärtää ja omaksua. Sanonta kuva kertoo enemmän kuin tuhat sanaa on jokaiselle tuttu lausahdus. Datasta on jopa sanottu sen olevan tulevaisuuden öljy. Ei siis ihme, että sen hyödyntämiseksi ovat syntyneet suuret markkinat joissa on kyse isoista rahoista. Suurten perinteisten Business Intelligence -toimijoiden rinnalle on syntynyt monia yrityskooltaan pienempiä tiedon visualisointiin keskittyneitä toimijoita. Näiden innovatiivisten uusien toimijoiden valttina on ollut ohjelmistojen helppokäyttöisyys, jossa ei välttämättä vaadita käyttäjältä ohjelmointitaitoja ja syvempää ymmärrystä tietomallien rakenteista. Isot toimijat ovat hieman jälkijunassa kehitelleet omista raskaammista ohjelmistoistaan vastaavanlaisia visualikykyisempiä ja helppokäyttöisiä versioita. Hienoa kuitenkin huomata, että markkinoille mahtuu kumpiakin toimijoita, sillä Gartenerin nelikentässä nämä innovatiiviset datavisualisoinnin ”uudet” toimijat ovat jo monta vuotta olleet näyttävästi esillä.

Raportoinnin ja analytiikan itsepalvelutyökalujen käyttäjät tutkivat itsenäisesti heille spesifioiduista datamalleista vastauksia kysymyksiinsä. Tiedän kokemuksesta että helppokäyttöisellä datavisualisointityökalulla työskentely on palkitsevaa ja kannustaa aidosti tutkimaan aineistoista erilaisia lineaarisia riippuvuuksia ja poikkeamia sekä esittämään löydöksiä monin eri visuaalisin vaihtoehdoin. Ne parhaat ja selkeimmät näkymät syntyvät kuitenkin noudattaen nöyrästi datavisualisoinnin parhaita käytäntöjä.

Grafiikan elementeistä värien käytössä piilee kuitenkin yksi merkittävä vaara. Maailman väestöstä 6 - 8 prosenttia (miehistä) ovat puna-vihervärisokeita (naisista 0,5 %) (Colblindor 2006). Näiden henkilöiden kohdalla voidaan datavisualisoinnin ymmärtämättömällä käytöllä epäonnistua visualisoinnin lopputuloksessa, kun esiin nostetut eri värielementein korostetut osa-alueet analyyseistä jäävät vaille huomioita. On syytä käyttää perusvärien kirjoa sekoittamatta esim. punaisen ja vihreän eri vivahteita samalle visuaaliselle raportille tai koosteraportille, kuten datavisualisoinnin kompastuskivet -kappaleessa todettiin.

On mielestäni perusteltua väittää datavisualisoinnin olevan merkittävässä asemassa myös tulevaisuudessa tiedon tehokkaimpana esitystapana, perustuen jo aiemmin esitettyihin hahmolakeihin sekä ihmisen luontaiseen kykyyn havainnoida kuvia ja värejä. Teknologian kehittyminen tuo meille lisää informaatiohaasteita. Big datan sekä IoT:n myötä meillä on jo nyt massiiviset määrät potentiaalista tietoa hyödynnettäväksi. Koneoppimisen, keinoälyn, algoritmien ja automatiikan avulla saadaan tiedon louhintaan ja sen jalostamiseen yhä tehokkaampia ja nopeampia keinoja. Tiedon hyödyntämisrajapinnassa ei voimakkaampaa ja tehokkaampaa tiedon välitystapaa ole kuin tiedon oikeaoppinen visualisointi.

Lähteet

Aerow 13.4.2017. Top 5 Self-Service Business Intelligence Tools - blog. Luettavissa: https://www.aerow.group/a17u1304/. Luettu: 12.5.2018.

Aunimo, L. 25.10.2017. Mitä on Big Data? Mitä eroa on Big Data – analytiikalla ja perinteisellä tilastollisella analyysillä? Luettavissa: https://www.bigdataresearch.fi/blog-4.-mita-eroa-on-big-data--analytiikalla-ja-perinteisella-tilastollisella-analyysilla. Luettu: 2.5.2018.

Colblindor 2006. Colorblind Population. Luettavissa: http://www.color-blindness.com/2006/04/28/colorblind-population/. Luettu: 14.5.2018.

Datatiede 2014. Analytiikan tasot. Luettavissa: http://www.datatiede.fi/analytiikan-tasot/. Luettu: 19.5.2018.

Enho, H. 30.1.2016. Power BI – kaikki mitä sinun tulee tietää aloitaessasi. Luettavissa: https://hexcelligent.fi/2016/01/30/power-bi-kaikki-mita-sinun-tulee-tietaa-aloittaessasi/. Luettu: 15.4.2018.

Holopainen 2016. Big data – nykytila ja tulevaisuuden mahdollisuudet. Luettavissa: http://www.tutuseura.fi/wp-content/uploads/2016/06/Futura2_16-BigData-avaus.pdf. Luettu: 6.5.2018.

Ilchenko, V. 28.2.2017. Informaticalla tietovirrat haltuun. Luettavissa: http://www.aureolis.com/big-data/informaticalla-tietovirrat-haltuun. Luettu: 22.5.2018.

IoTalents 2018. Examples of bad visualizations. Luettavissa: https://www.iotalents.com/forum/question/275253. Luettu: 14.5.2018.

Kanerva, J. 11.9.2016. Tiedon visualisointi-parhaat käytännöt. Luettavissa: https://infograafikko.fi/infografiikka/tiedon-visualisointi-parhaat-kaytannot/. Luettu: 28.4.2018.

Kolehmainen, A. 18.11.2011. Mitä eroa on big datalla ja perinteisellä datalla. Luettavissa: https://www.tivi.fi/CIO/2011-11-18/Mit%C3%A4-eroa-on-big-datalla-ja-perinteisell%C3%A4-datalla-3188167.html. Luettu: 27.4.2018.

Koski, J. 3.3.2015. Informaation visualisointi. Luettavissa: https://medium.com/@johanneskoski/informaation-visualisointi-e8615483680e. Luettu: 1.5.2018.

Laine, A. 18.2.2004. LuK-tutkielma: Hahmolait käytettävyyden parantajina. Luettavissa: http://www.mit.jyu.fi/opetus/opinnayte/LuK/Hahmolait/. Luettu: 29.4.2018.

Neisser, U. 1982. Kognitio ja todellisuus. Weilin+Göös. Espoo.

Niemelä 2018. Big Data, Data Mining – tiedonlouhinta. Luettavissa: https://www.hub.fi/ajankohtaista/blogi/70-big-data-data-mining-tiedonlouhinta. Luettu: 1.5.2018.

Niemijärvi, V. 2.4.2013. Ja maailman paras raportointiohjelmisto on. Luettavissa: http://www.louhia.fi/2013/04/02/ja-maailman-paras-raportointiohjelmisto-on/. Luettu: 20.4.2018.

Pengon 28.10.2015. Magic Quadrant – Mistä on kyse? Luettavissa: http://blogi.pengon.fi/magic-quadrant-mista-on-kyse. Luettu: 22.5.2018.

Qlik. 18.9.2014. Qlik visualisoi datan uudella Qlik Sense itsepalvelusovelluksella. Luettavissa: https://netprofile.fi/tiedotteet/qlik-visualisoi-datan-uudella-qlik-sense-itsepalvelusovelluksella/. Luettu: 16.5.2018.

Salo, I. 2013. Big data. Tiedon vallankumous. Docendo Oy. Jyväskylä.

Salo, I. 2014. Big data & Pilvipalvelut. Docendo Oy. Jyväskylä.

Sinkkonen, I. Kuoppala, H. Parkkinen, J. & Vastamäki, R. 2006. Käytettävyyden psykologia. Edita Prima Oy. Helsinki.

Taloussanomat 2011. Datan määrä räjähtää käsiin – IT osaaminen ei riitä. Luettavissa: https://www.is.fi/digitoday/art-2000001716471.html. Luettu: 3.5.2018.

Tilastokeskus 2018. Tietotekniikan käyttö yrityksissä. Luettavissa: http://www.stat.fi/til/icte/luo.html. Luettu: 12.5.2018.

Tilastokeskus 2018. PX-Web API. Luettavissa: https://www.stat.fi/org/avoindata/api.html. Luettu: 23.5.2018.

Tietokaira 2015. Tietovarastointi. Luettavissa: http://www.tietokaira.fi/tuotteet-ja-palvelut/tietovarastointi. Luettu: 15.4.2018.

Vakkuri, M. 20.6.2013. Big Data muuttaa maailmaa. Luettavissa: https://www.talouselama.fi/kumppaniblogit/big-data-muuttaa-maailmaa/6e3988d0-e07e-35ea-b52c-dc3e31a91394. Luettu: 22.5.2018.

Väisänen, A. 15.8.2017. Interaktiivinen visualisointi. Luettavissa: https://www.almpartners.fi/tag/interaktiivinen-visualisointi/. Luettu: 15.5.2018.

Liitteet

Liite 1. Magic Quadrant for Analytics and Business Intelligence Platforms

Magic Quadrant for Analytics and Business Intelligence Platforms

Published: 26 February 2018 **ID:** G00326555

Analyst(s):
Cindi Howson, Rita L. Sallam, James Laurence Richardson, Joao Tapadinhas, Carlie J. Idoine, Alys Woodward

Strategic Planning Assumptions

By 2020, augmented analytics — a paradigm that includes natural language query and narration, augmented data preparation, automated advanced analytics and visual-based data discovery capabilities — will be a dominant driver of new purchases of business intelligence, analytics and data science and machine learning platforms and of embedded analytics.

By 2020, the number of users of modern business intelligence and analytics platforms that are differentiated by augmented data discovery capabilities will grow at twice the rate — and deliver twice the business value — of those that are not.

By 2020, natural-language generation and artificial intelligence will be a standard feature of 90% of modern business intelligence platforms.

By 2020, 50% of analytical queries will be generated via search, natural-language processing or voice, or will be automatically generated.

By 2020, organizations that offer users access to a curated catalog of internal and external data will derive twice as much business value from analytics investments as those that do not.

Through 2020, the number of citizen data scientists will grow five times faster than the number of expert data scientists.

Tableau

Tableau is in the Leaders quadrant. Contributions to this position include its efforts to build product awareness globally; and its product roadmap, which includes NLP, augmented data preparation and discovery and agile data cataloging, among others. Strong market momentum in an increasingly competitive and price-sensitive market; ongoing product improvements; and excellent customer reference scores for customer experience and success also drive its position.

Qlik

Qlik's position in the Leaders quadrant is driven by progress on its roadmap for augmented analytics, improvements in marketing strategy, and ease of use. Its market execution is poorer than that of the other Leaders, largely due to its relatively low momentum, slightly lower product success, and its operations scores.

Microsoft

Microsoft is positioned in the Leaders quadrant again this year, with continued strong uptake of Power BI, and high levels of customer interest and adoption. Microsoft has clear and visionary product roadmap that includes vertical industry content.